Regularized Deep Neural Network for Post-Authorship Attribution

Abiodun Modupe

School of Computer Science and Applied Mathematics,
University of the Witwatersrand, Johannesburg, South Africa.

A thesis submitted to the Faculty of Science,
in fulfilment of the requirements for the degree of
Doctor of Philosophy in Computer Science.

Supervisors:
Professor Turgay Celik
Professor Vukosi Marivate

February 12, 2024

Abiodun Modupe
Regularized Deep Neural Network for Post-Authorship Attribution.

Copyright © University of the Witwatersrand, Johannesburg, South Africa.

SUPERVISORS:
Professor Turgay Celik[1]
Prof Vukosi Marivate[2]

SUPPORTED BY:
Council for Scientific and Industrial Research under the Department of Science and Technology in South Africa.

1 School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa.
2 Department of Computer Science, Head of Data Science, University of Pretoria, Lynnwood Road, Pretoria, South Africa.

Dedicated this thesis
to God Almighty my creator, my strong pillar, my source of inspiration, wisdom, knowledge and understanding. He has been the source of my strength throughout this program and on His wings only have I soared.

ACKNOWLEDGMENTS

The works presented in this thesis has been a life-changing experience and it would not have been possible without the support and guidance that I have received from many people. I, therefore, take this opportunity to extend my sincere gratitude and appreciation to all those who have made this PhD thesis possible.

Firstly, my highest appreciation goes to my Creator, the Lord, God Almighty for the gift of life; for his endless love, and for providing me strength and endurance to complete this work.

I am grateful to Professor Turgay for suggesting my topic and introducing me to the exciting fields, and Professor Vukosi Marivate for leaping over to natural language processing (NLP). Your consciousness and unconsciousness are the elements of how good research should be done. I acknowledge you all for your extraordinary patience, enthusiastic encouragement, and eternal cheerfulness; I witnessed the excellent leadership of an academic advisor. I learned everything I know about machine learning techniques from all the technical discussions we had together. I will forever value all the time, advice, inspiration, encouragement, and ideas he has contributed to my PhD pursuit. The joy and enthusiasm you both have for research are truly remarkable and have been a motivational factor throughout my PhD journey. The cry of my heart is that God should guide, protect and be with you all the days of your life. In my heart, you will be forever young.

I am extremely grateful to Prof Oludayo O Olugbara for his thoroughness, demeanour, and invaluable advice, continuous support, uncanny ability to identify and describe the most discrete items, and vast theoretical, technical, social and philosophical knowledge. Dear Prof Oludayo Olugbara, your immense knowledge and plentiful experience have encourage and take me this far in my academic life and daily life right from the first time I met you. Prof Oludayo Olugbara, I am deeply indebted to you contribution in this work. Thank you. Equally, my appreciation goes to Dr (Mrs) Olugbara for your love, encouragement, and unflinching support.

I am indebted to the Council for Scientific and Industrial Research - Department of Science and Technology of South Africa, under the Interbursary PhD bursary support programme for funding this research. I am also grateful to the Association for Deep learning Indaba, and Black in Artificial Intelligence for providing travel grants to the United States of America, Canada, Stellenbosch University, and Tunisia respectively.

I am also grateful to my classmates and cohort members, especially DSFI research group at the University of Pretoria and formal mates at WITS, for their editing help, late-night feedback sessions, and moral support. Thanks, should also go to the librarians, research assistants, and study participants both at WITS and UP, who impacted and inspired me.

Lastly and most importantly, I would like to thank my family, especially my wife for sharing in my dreams, support, love, and patience with me during those late nights spent in the research lab. My very lovely son, Oluwalana Oluwabamise Daniel Modupe, thank you. My appreciation also goes to my families, the Modupes, the Okitikans, and particularly Fatunla's family and Yemi, my friends for life, for making this journey exciting and fun.

DECLARATION

This thesis is submitted for the Doctor of Philosophy in Computer Science at the University of the Witwatersrand, Johannesburg. This work has not been submitted to any other University or for any other degree. I declare that this thesis is my original and unaided intellectual work.

Abiodun Modupe

February 12, 2024

ABSTRACT

Post-authorship attribution is the computational process of determining the legitimate author of an online text snippet, such as an email, blog, forum post, or chat log, by employing stylometric features. The process consists of analysing various linguistic and writing patterns, such as vocabulary, sentence structure, punctuation usage, and even the use of specific words or phrases. By comparing these features to a known set of writing pieces from potential authors, investigators can make educated hypotheses about the true authorship of a text snippet. Additionally, post-authorship attribution has applications in fields like forensic linguistics and cybersecurity, where determining the source of a text can be crucial for investigations or identifying potential threats. Furthermore, in a verification procedure to proactively uncover misogynistic, misandrist, xenophobic, and abusive posts on the internet or social networks, finding a suitable text representation to adequately symbolise and capture an author's distinctive writing from a computational linguistics perspective is typically known as a stylometric analysis. Additionally, most of the posts on social media or online are generally rife with ambiguous terminologies that could potentially compromise and influence the precision of the early proposed authorship attribution model. The majority of extracted stylistic elements in words are idioms, onomatopoeias, homophones, phonemes, synonyms, acronyms, anaphora, and polysemy, which are fundamentally difficult to interpret by most existing natural language processing (NLP) systems. These difficulties make it difficult to correctly identify the true author of a given text. Therefore, further advancements in NLP systems are necessary to effectively handle these complex linguistic elements and improve the accuracy of authorship attribution models.

In this thesis, we introduce a regularised deep neural network (RDNN) model to solve the challenges that come with figuring out who wrote what after the fact. The proposed method utilises a convolutional neural network, a bidirectional long short-term memory encoder, and a distributed highway network to effectively address the post-authorship attribution problem. The neural network was utilised to generate lexical stylometric features, which were then fed into the bidirectional encoder to produce a syntactic feature vector representation. The feature vector was then fed into the distributed high-speed networks for regularisation to reduce network generalisation errors. The regularised feature vector was then given to the bidirectional decoder to learn the author's writing style. The feature classification layer is made up of a fully connected network and a SoftMax function for prediction. The RDNN method outperformed the existing state-of-the-art methods in terms of accuracy, precision, and recall on the majority of the benchmark datasets. These results highlight the potential of the proposed method to significantly improve classification performance in various domains.

Again, the introduction of an interactive system to visualise the performance of the proposed method would further enhance its usability and effectiveness in quantifying the contribution of the author's writing characteristics in both online text snippets and literary documents. It is useful in processing the evidence that is needed to support claims or draw conclusions about the author's writing style or intent during the pre-trial investigation by the law enforcement agent in the court of law. The incorporation of this method into the pre-trial stage greatly strengthens the credibility and validity of the findings presented in court and has the potential to revolutionise the field of authorship attribution and enhance the accuracy of forensic investigations. Furthermore, it ensures a fair and just legal process for all parties involved by providing concrete evidence to support or challenge claims.

We are also aware of the limitations of the proposed methods and recognise the need for additional research to overcome these constraints and improve the overall reliability and

applicability of post-authorship attribution of online text snippets and literary documents for forensic investigations. Even though the proposed methods have revealed some unusual differences in author writing style, such as how influential writers, regular people, and suspected authors use language, the evidence from the results with the features extracted from the texts has shown promise for identifying authorship patterns and aiding in forensic analyses. However, much work remains to be done to validate the methodologies' usefulness and dependability as effective authorship attribution procedures. Further research is needed to determine the extent to which external factors, such as the context in which the text was written or the author's emotional state, may impact the identified authorship patterns. Additionally, it is crucial to establish a comprehensive dataset that includes a diverse range of authors and writing styles to ensure the generalizability of the findings and enhance the reliability of forensic analyses.

Furthermore, the dataset used in this thesis does not include a diverse variety of authors and writing styles, such as impostors attempting to impersonate another author, which limits the generalizability of the conclusions and undermines the credibility of forensic analysis. More studies can be conducted to broaden the proposed strategy for detecting and distinguishing impostors' writing styles from those of authentic authors when committing crimes on both online and literary documents. It is conceivable for numerous criminals to collaborate to perpetrate a crime, which could aid in improving the proposed methods for detecting the existence of multiple impostors or the contribution of each criminal writing style based on the person or individual they are attempting to mimic. The likelihood of numerous offenders working together complicates the investigation and necessitates advanced procedures for identifying their individual contributions, as well as both authentic and manufactured impostor contents within the text. This is especially difficult on social media, where fake accounts and anonymous profiles can make it difficult to determine the true identity of those involved, which can come from a variety of sources, including text, WhatsApps, chat images, videos, and so on, and can lead to the spread of misinformation and manipulation. As a result, promoting a hybrid approach that goes beyond text as evidence could help address some of the concerns raised above. For example, integrating audio and visual data may provide a more complete perspective of the scenario. As a result, such an approach exacerbates the restrictions indicated in the distribution of data and may necessitate more storage and analytical resources. However, it can also lead to a more accurate and nuanced analysis of the situation.

CONTENTS

I Introduction

1 Introduction 3
- 1.1 Problem Description . . . 6
  - 1.1.1 Cybercrime investigation . . . 6
  - 1.1.2 Social network analysis and political socialisation . . . 7
  - 1.1.3 Literary science and education . . . 8
- 1.2 Problem statement . . . 9
- 1.3 Research Context . . . 10
  - 1.3.1 Research challenge . . . 10
  - 1.3.2 Research Questions . . . 11
- 1.4 Technical contributions . . . 12
- 1.5 Study outline . . . 12

II Literature Review

2 Background and Related Work 17
- 2.1 Social media . . . 17
  - 2.1.1 Twitter . . . 18
  - 2.1.2 Characteristics of Twitter . . . 21
  - 2.1.3 Cybercriminal activity on social media via tweets . . . 23
- 2.2 Feature metrics . . . 24
  - 2.2.1 Style . . . 25
  - 2.2.2 Lexical . . . 28
  - 2.2.3 Syntactic . . . 28
  - 2.2.4 N-gram . . . 30
  - 2.2.5 Semantic metrics . . . 33
  - 2.2.6 Application-specific . . . 34
  - 2.2.7 Multiple feature metrics combinations . . . 35
  - 2.2.8 Feature metrics Summary . . . 36
- 2.3 Authorship analysis . . . 36
  - 2.3.1 Authorship attribution . . . 37
  - 2.3.2 Authorship verification . . . 40
  - 2.3.3 Writer profiling . . . 41
  - 2.3.4 Style change detection . . . 42
- 2.4 Limitations of the existing authorship method . . . 43
  - 2.4.1 Interpretation the results . . . 43
  - 2.4.2 Small number of writers . . . 43
  - 2.4.3 Training data size . . . 44
- 2.5 Chapter Summary . . . 46

III Automating Exploratory Statistical Analysis with Machine learning

3 Automating Exploratory Statistical Analysis with Machine learning 49
- 3.1 Introduction . . . 49
- 3.2 Authorship Attribution Dataset . . . 51
  - 3.2.1 CCAT50 dataset . . . 52
  - 3.2.2 IMDb62 dataset . . . 52
  - 3.2.3 Blog dataset . . . 52
  - 3.2.4 Twiiter dataset . . . 53
- 3.3 Exploratory statistical analysis and visualisations . . . 53
  - 3.3.1 Histograms . . . 54
  - 3.3.2 Quantile-normal plots . . . 54

3.3.3 Correlation and covariance . . . 55
3.3.4 Sentiment factors . . . 57
3.3.5 Interactive tools for ESA visualization . . . 58
3.4 Chapter Summary . . . 65

IV Deep regularized neural networks for PAA tasks
4 Regularized Deep Neural Network for Post Authorship Attribution 71
4.1 Introduction . . . 71
4.2 Proposed Method . . . 71
4.2.1 CNN Character-Level Layer . . . 72
4.2.2 Distributed Highway Network with Bidirectional Long Short-Term Memory . . . 74
4.2.3 Feature-Classification Layer . . . 75
4.3 Experimental Results . . . 76
4.3.1 Datasets . . . 76
4.3.2 Parameter Settings . . . 76
4.3.3 Result . . . 77
4.3.4 Result Visualization . . . 81
4.4 Discussion . . . 85
4.5 Chapter Summary . . . 87

V Application of Sub-word embedding for PAA tasks
5 Authorship Identification by Integrating Bidirectional Long short-term memory with Subword Embedding 91
5.1 Introduction . . . 91
5.2 Related Work . . . 92
5.3 Proposed method . . . 94
5.3.1 Text Representation . . . 94
5.3.2 Feature Extraction . . . 95
5.4 Experiment and results . . . 98
5.4.1 Datasets . . . 98
5.4.2 Hyperparameters tuning and training . . . 98
5.4.3 Results . . . 99
5.5 Discussion . . . 100
5.6 Chapter Summary . . . 101

VI Post-Authorship Attribution Using Attention-based Bidirectional Gated Recurrent Units with Convolution
6 Post-Authorship Attribution Using Attention-based Bidirectional Gated Recurrent Units with Convolution 105
6.1 Introduction . . . 105
6.2 Related Work . . . 107
6.3 Preliminaries . . . 111
6.3.1 Input layer . . . 112
6.3.2 Embedding layer . . . 112
6.3.3 Feature extraction and representation layer . . . 114
6.4 Proposed System . . . 117
6.5 Evaluation and results . . . 119
6.5.1 Parameter settings . . . 119
6.5.2 Evaluation criteria . . . 120
6.5.3 Results . . . 121
6.5.4 Result Visualization . . . 126
6.5.5 Chapter Summary . . . 128

VII Conclusions & Future work
7 Conclusion and future work 133
7.1 Summary . . . 133
7.2 Future Work . . . 135
Bibliography 232

LIST OF FIGURES

Figure 1 Layout platform of the third-party interface to access twitter [Chen *et al.* 2022]. . . . 20
Figure 2 Number of followings and followers. . . . 22
Figure 3 The total number of followers and tweets per user. . . . 22
Figure 4 A online text snippets from hacker group. . . . 24
Figure 5 Authorship analysis tasks. . . . 38
Figure 6 A snippet from CCAT50 datasets . . . 52
Figure 7 A snippet from IMDb62 datasets . . . 52
Figure 8 A snippet from Blog datasets . . . 53
Figure 9 Histogram and CDF on the CCAT50 datasets. . . . 54
Figure 10 Histogram and CDF on the Twitter datasets. . . . 55
Figure 11 A quantile-normal plot. . . . 55
Figure 12 Sample covariance heatmap for Twitter dataset. . . . 56
Figure 13 Unique and average distribution in the CCAT50 datasets. . . . 57
Figure 14 Sample covariance heatmap for CCAT50 dataset. . . . 57
Figure 15 Unique and average distribution in the Twitter datasets. . . . 58
Figure 16 Sentiment factors. . . . 59
Figure 17 Phrases score extracted using PyTextRank algorithm [Mihalcea and Tarau 2004] for Twitter dataset. . . . 61
Figure 18 Phrases extracted using PyTextRank for Twitter dataset. . . . 62
Figure 19 Phrase-machine [Handler *et al.* 2016] for Twitter dataset. . . . 63
Figure 20 Example of Phrase-machine with positive sentiment terms in Twitter dataset. . . . 63
Figure 21 Example of Phrase-machine with negative sentiment terms in Twitter dataset. . . . 64
Figure 22 Example of topics and categories visualization over a sample text in Twitter dataset. . . . 66
Figure 23 Visualizing query based on category different. . . . 67
Figure 24 The architecture of the RDNN PAA system. . . . 73
Figure 25 Performance accuracy and loss curves over epoch on the CCAT50 datasets. 79
Figure 26 Performance accuracy and loss curves over epoch on the IMDB62 datasets. 79
Figure 27 Performance accuracy and loss curves over epoch on the Blog50 datasets. 79
Figure 28 Performance accuracy and loss curves over epoch on the Twitter50 datasets. . . . 80
Figure 29 Performance accuracy on CCAT50 and IMDB62 dataset compared to SOTA methods with varying numbers of authors. . . . 80
Figure 30 Explanation of individual prediction of the proposed method on the CCAT50 dataset using LIME. . . . 82
Figure 31 Explanation of individual prediction of the proposed method on the IMDB62 dataset using LIME. . . . 83
Figure 32 Explanation of individual prediction of the proposed method on the Twitter50 dataset using LIME. . . . 84
Figure 33 Explanation of evaluation results based on the SHAP tree explainer for CCAT50 dataset. . . . 85

Figure 34 Explanation of evaluation results based on the SHAP tree explainer for IMBDB62 dataset. 85
Figure 35 SHAP summary plots for CCAT50 and IMDB62 dataset 86
Figure 36 BLSTM-CNN model with subword information 94
Figure 37 The accuracy and convergence curve over iterations on all the datasets 100
Figure 38 The accuracy and convergence curve over iterations on Twitters dataset. 101
Figure 39 ABC model 112
Figure 40 CBOW and the Skip-gram model for word embedding representations. 114
Figure 41 The attention mechanisms are fitted with bidirectional LSTM and GRU networks. 116
Figure 42 Convolutional layer to extract local feature. 117
Figure 43 Quantile-quantile statistical distribution of the datasets 120
Figure 44 Performance accuracy and loss curves over epoch on the CCAT50 datasets. 123
Figure 45 Performance accuracy and loss curves over epoch on the Twitter50 datasets. 125
Figure 46 ROC curve for IMDB62, Blog and Twitter datasets with 10 authors. . . 125
Figure 47 Confusion matrix for IMDb62 datasets with 10 & 62 authors. 126
Figure 48 Confusion matrix for Blog datasets with 10 & 50 authors. 126
Figure 49 Confusion matrix for Twitter datasets with 10 & 50 authors. 127
Figure 50 Explanation of individual prediction of the proposed ABC model on the IMDb50 dataset using LIME. 127
Figure 51 Explanation of individual prediction of the proposed ABC model on the Blog 50 dataset using LIME. 128
Figure 52 Explanation of individual prediction of the proposed ABC model on the Twitter50 dataset using LIME. 128

LIST OF TABLES

Table 2 Basic features metric for authorship analysis [Neal *et al.* 2017a; Sari 2018]. 37
Table 3 Statistics of the datasets 51
Table 4 TextRank algorithm scores for Twitter datasets. 61
Table 5 The setting of hyperparameters in our model on different datasets . . . 76
Table 6 Performance ACC and F1 of the RDNN method and SOTA methods for PAA sub-task. 78
Table 7 Performance variance through repeated runs on the Blog50 datasets . . 80
Table 8 Performance variance through repeated runs on the Twitter50 datasets. 81
Table 9 BPE example given raw input text. 96
Table 10 Performance comparison and accuracy scores on four mainstream AA datasets. 99
Table 11 Performance comparison of CNN and BretAA model on all the datasets. 100
Table 12 Comparison of related studies based on common characteristics 111
Table 13 Comparison of the results obtained on three datasets for 50 authors . . 122
Table 14 Results for 20 authors 124
Table 15 Results for 10 authors 124

AA	Authorship Attribution
AFF	Advanced fee fraud
ANN	Artificial neural networks
AP	Authors profiling
API	Application Programming Interface (API)
AV	Authorship verification
BLSTM	Bidirectional long short-term memory
BOW	Bag-of-words
CD	Cardinal numbers
CDF	Cumulative Distribution Function
CLEF	Conference and Labs of the Evaluation Forum
CNN	Convolutional neural network
DT	Determiners
EER	Equal error rate
FFNN	Feed-forward neural network
FW	Function words
IGE	Information gain and entropy
IP	Internet Protocol
IR	Information retrieval
ISOT	In-formation security and object technology
JJ	Adjective
LDA	Latent Dirichlet Allocation
LIME	Local Interpretable Model-agnostic Explanations
LIWC	Linguistic inquiry and word count
LTS	Long text samples
MCSAN	Multichannel Self-Attention Network
MLA	Machine Learning Algorithm
ML	Machine learning
NLP	Natural language processing
NLTK	Natural language toolkit
NNP	Proper Nouns
NNS	Common noun plural form
NRC	National research council
OTS	Online text snippets
PAA	Post-authorship attribution
PAN	Plagiarism detection, author identification and author profiling
PCA	Principal component analysis
POS	Part-of-speech
RDNN	Regularised deep neural networks

RF	Random Forest
RNN	Recurrent neural networks
ROC-Curve	Receiver Operating Characteristic Curve
SBD	Style breach detection
SCD	Style change detection
SC	Source code
SHAP	Shapley Additive Explanations
SMS	Short Message services
SMO	Sequential minimal optimisation
SOTA	State-of-the-art
SVM	Support Vector Machine
TC	Telegraph columnist
VBG	Verb gerund or present
VBP	Verb non-3rd person singular present form
VADER	Valence aware dictionary for sentiment reasoning
VP	Verb base form
WP	Writing profiling

PUBLICATIONS

Some portions of this thesis have been published in the following articles. These portions include figures, algorithms, equations, and descriptions of concepts. These papers are:

[1] A. Modupe, T. Celik, V. Marivate and M. Diale, *Semi-supervised probabilistics approach for normalising informal short text messages*. In the proceedings of IEEE Conference on ICT and Society (ICTAS), pp 1-8, ISBN:978-1-4673-8996-9, March 8th to 10th, 2017. Durban, South Africa. URL: https://10.1109/ICTAS.2017.7920659.

[2] Modupe A, Celik T, Marivate V, Olugbara OO. .Post-Authorship Attribution Using Regularized Deep Neural Network. *Applied Sciences*. (2022);12(15):7518. DOI: https://doi.org/10.3390/app12157518.

[3] Modupe A, Celik T, Marivate V, Olugbara OO (2023). *Integrating Bidirectional Long Short-Term Memory with Subword Embedding for Authorship Attribution*. In the proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), URL: https://ieeesmc2023.org/. *Submitted–February 2023: United State,Hawaii*.

[4] Modupe A, Celik T, Marivate V, Olugbara OO (2023). *Post-Authorship Attribution Using Attention-based Bidirectional Gated Recurrent Units with Convolution*. In the IEEE Transactions on Computational Social Systems, URL: https://www.ieeesmc.org/publications/transactions-on-computational-social-systems/. *Submitted–April 2023*.

Part I

INTRODUCTION

The problem of finding the legitimate author of the given online text snippet based on the reflected writing styles is known as post-authorship attribution (PAA). PAA is a computationally driven method that employs multiple linguistic features with machine learning methods to characterize and identify anonymous writers of texts for forensics applications. The recent advances in internet technology coupled with the blossoming of electronic writing, such as online text snippets in modern society have made PAA for analysis and identification necessary in modern society. The application of PAA becomes apparent in a situation when criminals send unlawful email threats or post tweets via Twitter pages, blogs, and online forums for financial reward. PAA, through the use of authorship attribution, can be divided into two problems. The first *closed-set attribution* (Section 2.3). In this problem, the challenge is to determine the genuine author of anonymous or questioned texts used in crimes among the candidates on the list (e. g. where the set of authors is known). The second problem is *open-set attribution* (Section 2.3.2), which is also known as the *author verification* problem where the genuine author may not be included in the list of candidates [Koppel *et al.* 2009; Stamatatos 2009b].

By combining statistical style and content-based features with a machine learning algorithm, it has been extremely difficult to demonstrate how similar two documents are and figure out who wrote them. This is uncommon and depends on the scenario or dataset. The aforementioned methods fail to learn appropriate writing styles when texts from known authors in the training set differ from those from disputed authors in the test set in terms of topic or genre. Because the writing styles of different topics or genres differ (for example, most online text snippets or texts have more concise sentences than prose), there are stringent requirements for accurately capturing the stylistic properties of texts related to the personal styles of authors if the training data do not contain information related to the disputed documents' topic or genre. The final challenge is to define appropriate stylometric metrics that capture the author's writing style and influence the choice of an appropriate classification method, which should be used as language-based evidence during the pre-trial investigation to determine the anonymous or questioned texts used for crimes. Although a variety of features and methods have been used to accomplish this task, as the number of authors increased, most of the old methods became less valuable, and they lacked transparency and interpretation about the author's writing style because the input text was usually simple snippets with short sentences and many words that made them difficult to understand, such as idioms, onomatopoeias, and homophones. This thesis investigates various authorship attribution methods and proposes new methods for learning an author's writing style, as well as the influence of feature vectors on PAA systems across multiple datasets. This is useful during an evidence-based forensic investigation.

This part contains two chapters. Chapter 1 provides an introductory context for this research. Chapter 2 provides a discussion on how to research topics that have developed in authorship attribution in forensic analysis and provides an overview of fundamental feature measurements and analysis methods, respectively. After that, we outline the main problems in authorship attribution research and present our methods to resolve the heightened challenges.

1 INTRODUCTION

Language is a natural human endowment, varying for the distinction among cultures. As a cultural heritage, language is a carrier of values and knowledge, often used in the practice and transmission of intangible cultural heritage. The spoken word in each language is vital in enacting and conveying intangible heritage, particularly in oral traditions and expressions, songs, and rituals. Even though languages are used primarily for communication, they also help with the equally important identity purpose [Adegbola 2021]. Cultures and communities may, therefore, be defined by geography and other spatial realities, signifying spoken words in any language as a crucial intangible heritage. Language forms a foundation for how an individual is perceived or identified. Humans are social animals; therefore, the words from the language reveal terms and writing styles, articulating thought in language that could characterize an individual, entity, location, or nation around an issue in a community. The intrinsic linguistic structure of the human mind guides this action [Kramsch 2004].

With the pervasiveness of portable computing devices and the Internet globalization and social networking sites [Ding *et al.* 2017], the language of information is constructed day-to-day. Profitable attempts are ensured by scripting expository, descriptive, persuasive, or narrative essays [Jeffrey 2016]. For example, expository writing is a common type where the author attempts to explain a concept and impart information to a broader audience. Expository writing does not include the author's opinions but focuses on accepted facts about a topic, including statistics or evidence. Descriptive writing is often established in fiction, though it can also appear in nonfiction (memoirs, first-hand accounts of events, or travel guides). When an author writes in a descriptive style, they attempt to paint a picture in words of a person, place, or object for a larger audience.

Metaphors or other literary devices might describe the author's impressions about what they hear, see, smell, taste, or touch (five senses). Persuasive writing attempts to convince the audience based on a position, idea, or belief (the dominant writing style in academic reports). Persuasive writing contains the author's opinions, biases, justifications, and reasons, provided as evidence of their position on a concern or challenge in a community. Narrative writing is used in most short or longer pieces, such as fiction or nonfiction. When writers use a narrative style, they are not only attempting to impart information, but they are also attempting to construct and communicate a story, complete with characters, conflict, and settings [Jeffrey 2016].

Each genre of language or writing in a text or document is used for a specific purpose [Paltridge 2012]. For example, online text snippets, such as Twitter[1], Facebook, SMS, email, blogging, Instagram, and WhatsApp, are constantly generated from varying sources to deliver valuable, entertaining content to people anywhere [Chen *et al.* 2020]; however, single online text snippets and documents may include more than one writing style attributed to an author or multiple writers. This information is frequently posted on social networks each millisecond in ideas, stories, opinions, sentiments, and other expression modes [Chen *et al.* 2018].

Twitter is a prominent microblogging service that allows users to communicate with one another and share information regardless of their physical location [Toraman *et al.* 2022].

1 also known as "**X**" [Stokel-Walker 2023] as discussed in Chapter 1, was founded in March 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams[2] and launched in July of that year [Kamberg 2012].

It has grown in popularity and become an indispensable tool for social networking and communication. It was established in October 2006 [Java *et al.* 2007]. The platform's popularity increased steadily over the years, but during the COVID-19 lockdown [Alshamrani *et al.* 2020], it grew to over 330 million monthly active users, making it the world's most popular social media platform [Ziwei and Chua 2019; Gharavi *et al.* 2020]. Twitter[3] is a famous platform with an enormous volume of dynamic users and a high revenue base. Twitter posts are widely called "Tweets". A blog is a web page that a blogger maintains to transform user communication and overcome the unidirectionality of standard online communication [Badea 2014]. Blogs or discussion forums can be an important source of information to describe customers' perspectives, such as product reviews, rating (dissatisfying or unsatisfactory) shopping behaviour, and post-purchase behaviour to enhance decision-making [Leeflang *et al.* 2014]; 600 million online blogs were produced and perused by 77% of users in July 2021 [Lin 2021]. In general, the proliferation of online activity has given rise to a worldwide pandemic, which has led to more deaths during COVID-19 through misinformation and fake news while also increasing the amount of illegal behavior. This has highlighted the need to develop critical analytical methods and media literacy proficiencies to combat the spread of misinformation and to ensure that perpetrators' activities can be expressed in a natural language in order to gain societal credibility, which has been lacking in recent years. It is crucial for individuals to be able to distinguish between reliable sources and false information.

The use of email is essential in organisational communication but can also serve as a pertinent source of work interruption and an impediment to well-being; however, it remains the most ubiquitous medium for organisational communication [Wijngaards *et al.* 2022]. As a result, 59% of working people read their work emails outside of normal business hours [Minnen *et al.* 2021]. However, we refer to after-hours e-mailing as reading and responding to business-related emails in lieu of conventional work hours. This behaviour has both negative and positive consequences for the employee's well-being. After-hours e-mailing can interfere with daily employee recovery, leading to negative effects such as exhaustion and diminished vigour [Minnen *et al.* 2021]. After-hours e-mailing can improve employee well-being by allowing them to complete an abandoned job and prepare workers for the next day [Minnen *et al.* 2021]. It can also perpetuate negative activation, e. g.affective rumination, increased stress, and burnout [Chen and Karahanna 2018; Minnen *et al.* 2021]. Moreover, study after study has shown that the constant availability and expectation of responding to after-hours communications blur the lines between work and personal life, resulting in a deficiency of work-life equilibrium [Minnen *et al.* 2021]. Additionally, being unable to separate work during non-working hours has a negative impact on relationships and the overall well-being of the individual outside of work, and people do this, especially on social media. Additionally, being unable to differentiate between work and non-working hours has a negative impact on relationships and an individual's overall well-being outside of work, which is something that many individuals do, particularly on social media. The constant connectivity of wanting to check on popular issues on social media and engage in conversation with others increases stress and burnout, causing people to abandon their jobs and focus on trivial thoughts and sensations. Ultimately, this reduces productivity and overall well-being [Burke *et al.* 2010; Verduyn *et al.* 2017].

We now live in a society where the digital world constantly diverts our attention and overwhelms us, which has a negative impact on our mental health and ability to form real human connections [Wadhwa and Palvia 2018]. This is called an era of "infobesity" in which people practise the language through words and phrases freely shared on social media [Ding *et al.* 2017], and it has led to a decline in the quality of real-life interactions as individuals rely heavily on online platforms for communication [Ryan *et al.* 2017; Arampatzi *et al.* 2018]. Any targeted

3 In recent years, the new owner has changed the bird's logo to "X" [Stokel-Walker 2023]

audience, whether online or offline, can access conversations, and most of the aforementioned posts are misused or misinterpreted out of context, leading to disagreements and disputes and helping to promote nocuous activities such as pirated software, terrorist communications, child pornography, grossly belligerent, harassing, abusive, obscene, menacing, corrosive, and antediluvian messages [Kešelj *et al.* 2003; Frantzeskou *et al.* 2006b; Escalante *et al.* 2011; Neal *et al.* 2017a]. Unfortunately, these have severe consequences for society, including the proliferation of harmful ideologies and the erosion of trust within communities. During COVID in 2019, for example, tweets or online text snippets sent via Twitter or social media were reported or purposely posted to mislead the general public about measures to take to avert the spread of the pandemic, and these increased the death rate all over the world, especially in the United States of America [Gabarron *et al.* 2021]. These misleading messages caused confusion and hindered the effectiveness of public health efforts, which led to further challenges in controlling the outbreak [Bin Naeem and Kamel Boulos 2021], eventually resulting in a higher number of cases and fatalities [Dubey *et al.* 2020; Chen *et al.* 2023].

Text is a common element for posting most of the social media messages with a unique concealment feature [Alonso-Fernandez *et al.* 2021]. The delinquents usually mask their identities and their writing style across multiple networks to circumvent authorship attribution tools [Ding *et al.* 2015]. An obfuscation process, where the perperator alters the textual context to prevent the authorship attribution tool from assigning incorrect authorship to the text and concealing the user identity without sacrificing the privacy of the user [Altakrori *et al.* 2018; Mahmood *et al.* 2020]. Moreover, it is also a common phenomenon for a user to masquerade as the original Internet protocol (IP)[4] addresses and hashtags, mentions people in texts, and reposts other authors' scripts [Zheng *et al.* 2006; Rocha *et al.* 2016; Theophilo *et al.* 2021]. The address of a genuine sender can be impersonated through an unidentifiable server to dispense offensive text messages across diverse pseudonymous channels [Zheng *et al.* 2006; Rocha *et al.* 2016; Theophilo *et al.* 2021]. The anonymity of social media posts imposes unique challenges and risks of tracing an offender's identity without compromising cyberspace privacy [Yu and Ginsberg 2018].

Stylometric techniques are fundamental to understanding the linguistic characteristics of authors in a selected literary text or document. They have traditionally been used in communicational linguistics and authorship attribution [Koppel *et al.* 2009; Stamatatos 2009b;b]. This method can also be used for forensic linguistics, which looks at the writing style of an anonymous poster to find out who wrote it [Rocha *et al.* 2016; Venčkauskas *et al.* 2015; Alonso-Fernandez *et al.* 2021], and for plagiarism detection, which compares the writing styles of different documents to see if one has been copied from another. Although stylometric techniques allow academics and investigators to use authorship attribution to get insight into the motivations, opinions, and potential biases of individuals behind online content by correctly identifying the source of anonymous posts. However, most works in this area employ extensive feature engineering to understand the author's identity and social-linguistic traits [Koppel *et al.* 2009; Stamatatos 2009b;b; Rocha *et al.* 2016; Venčkauskas *et al.* 2015; Alonso-Fernandez *et al.* 2021]. The aforementioned are time-consuming and labour-intensive tasks that require advanced techniques and extensive resources, e. g.to generate distinctive representation for every new task [Ding *et al.* 2017; Vosoughi *et al.* 2016; Zheng and Jin 2022]. Due to limited amounts of data or documents, as well as the fact that each training set does not contain a reasonable number of words. Also, the stylometry features are feature engineering variables and cannot be generalised to new tasks or explained [Ramnial *et al.* 2016a; Zheng and Jin 2022]. Therefore, the accuracy of the proposed model cannot be guaranteed [Sari 2018; Quiring *et al.* 2019]. Therefore, post-authorship attribution (PAA) is expanded to turn a lot of web content into meaningful representations that can pull out linguistic intuitions about the author's writ-

4 A device attached to a computer network with an easily recognisable number.

ing style. This includes social media text that is made up of online text snippets, blogs, and other types of web content [Sari 2018; Quiring *et al.* 2019].

The quantitative representation of the underlying features can be observed as a fingerprint to quantify an author's writing style [Kalgutkar *et al.* 2019]. The authorship attribution process has multifarious applications in the disciplines of computational linguistics [Hirschberg and Manning 2015], information retrieval [Stamatatos 2009b; Burrows *et al.* 2009], data mining [Stein *et al.* 2011], natural language processing [Stamatatos 2009b; Pillay and Solorio 2010; Juola and others 2008], and personal electronic mail [Iqbal *et al.* 2008a; 2013; Ding *et al.* 2015]. It can detect plagiarism in an article. This is a scenario where, for instance, an editor of a journal identifies whether a manuscript submitted for publication was written by someone else or published elsewhere [Stamatatos 2009b; Stein *et al.* 2011; Foltỳnek *et al.* 2019].

Authorship analysis or attribution (AA) is highly profitable in marketing for identifying potential customers through their reviews [Weren *et al.* 2014], understanding audiences' political preferences, and identifying political communication networks [Golbeck and Hansen 2011; 2014]. In many fields, like literary science, authorship attribution (AA) has a lot of prospects. For instance, it can be used to find the author of a disputed or anonymous text document [Klarreich 2003; Hoover 2004; Fabien *et al.* 2020b], identify anonymous phishing emails [Lambers and Veenman 2009; Iqbal *et al.* 2013], and uncover advanced fee fraud (AFF) [Modupe *et al.* 2011a; 2012a]. In general, this demonstrates that language was an essential cue to prove that Robert Galbraith was actually J.K. Rowling's pen name [Juola 2015]. Therefore, in many cases, the textual content of an online text snippet is the only evidence available and can be proven as substantial evidence in a Court of Law during a pre-trail investigation [Rudman 1997; Venčkauskas *et al.* 2015]. Researchers like Zimmer [2013]; Juola [2021] have defended the use of text evidence gathered from the crime scene in legal proceedings. These scholars argue that textual evidence can provide valuable insights and support in criminal investigations and court cases.

Van Halteren *et al.* [2005] implemented a human stylome to determine the writing style to identify the author. Similarly, Juola and others [2008] used the term authorial fingerprint to learn a pattern of language or writing style in provided texts. The aim was to use a person's writing style as a unique identification strategy, such as a fingerprint [Kalgutkar *et al.* 2019]; however, it has yet to be determined whether specific writing style elements or features could identify a person or an anonymous user in an online text snippet used in a crime such as a fingerprint [Malyutov 2005]. These writing style features must be consistent regardless of varying aspects, such as the writing time and document genre. A principal authorship attribution challenge is realizing a suitable writing style, characterizing a writer [Rocha *et al.* 2016; Alonso-Fernandez *et al.* 2021; Theophilo *et al.* 2021; Suman *et al.* 2021].

1.1 PROBLEM DESCRIPTION

With the accelerated Internet evolution, the number of online text snippets also increased. The anonymous nature of the abundant data provided rise to illicit possibilities [Sharma *et al.* 2018]. This is achieved by examining the reflected linguistic footprints in the text to uncover sociolinguistic characteristics to identify the authors of anonymous online text snippets. Identification and verification received attention in domains described in the subsequent sections.

1.1.1 *Cybercrime investigation*

The Internet or social media proliferation provides an anonymous platform for computer-mediated adversary activities, such as phishing scams, spamming, ransom messages, harassment, money laundering, and illegal material distribution. The network origin, such as

Internet Protocol (IP) addresses, can be quickly repudiated [Zheng *et al.* 2006; Rocha *et al.* 2016; Ding *et al.* 2017]. Several tactics have been proposed for determining the authorship of a specific LTS or OTS uploaded on social media, including the use of stylometric traits as evidence and the judgements of legal experts in court cases (koppel2006authorship, stolerman2014breaking,yang2014authorship). These tactics are designed to establish the legitimacy and validity of the authorship claims, which can be crucial in determining the credibility of information shared on social media. In addition to these tactics, another approach for determining authorship on social media involves analysing metadata associated with the post, such as the time and location of the upload. This can provide valuable insights into the identity of the author and help verify their claims. Furthermore, advancements in machine learning algorithms have also been utilised to analyse linguistic patterns and writing style, further strengthening the accuracy of authorship attribution techniques. These algorithms can detect subtle nuances in language use and writing style that can help identify the author of a post with a high level of confidence. Additionally, collaboration between researchers and social media platforms can lead to the development of more sophisticated tools and techniques for authorship attribution, ensuring the reliability of information shared on these platforms. In a well-known case in the United Kingdom and during the pre-trial investigation, the linguistic experts, through lawyers, could prove that an SMS series (questioned or anonymous texts used in crimes) from Danielle Jones' phone was by her uncle. Danielle disappeared on 18 June 2001 [Mohan *et al.* 2010; Grant 2020].

The forensic analyzes of online text snippets or documents for approaching the anonymity problem is a day-to-day modern challenge towards solving security. These include obfuscation [Brennan *et al.* 2012; Kalgutkar *et al.* 2019], multilingual domain challenges [Bogdanova and Lazaridou 2014], and privacy concerns in healthcare systems [Newaz *et al.* 2021]. Most methods used to identify the author of a provided text employed natural language processing (NLP). NLP is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language. It signifies how computers can process and analyze substantial natural language data. This involves analyzing the writing styles and stylometric features of the document content. PAA aims to classify provided texts based on the stylistic choices of their authors. Most proposed methods consider stylometric features only, such as lexical, syntactic, and semantic. This limitation complicates capturing the syntactic or semantic text features entirely to understand the author's writing style better.

1.1.2 *Social network analysis and political socialisation*

Inferring the hidden stylometric attributes of users could help to understand the perspectives and related discussions of the fans and followers more utterly. Topics include politics, business, and social wel-being [Ding *et al.* 2017]. Simaki *et al.* [2016] employed text mining, sociolinguistic-based, and content-related text features to identify men and women of various ages based on their online Twitter snippets. Conover *et al.* [2011b] attempted to understand the audiences on social networks using stylometric features, which helps in planning political strategy while correcting bias for predicting election results [Wong *et al.* 2013]. The social media revolution allowed people to share group individualism if a writer's style expresses the selection of hate words within the structure of the provided text or sentence.

Etaywe [2018] investigates the possibility of determining patterns in language used by Jordanian male and female authors by employing linguistic inquiry and word count (LIWC) to analyze the stylometric features of online texts. Communication digitalization is no longer a phenomenon confined to Western countries. Social media is almost dominating the Arab regions; 79% of people in the Middle East use social media or direct messaging at least once a day. This is misinformation with the harmful intention to cause satire, parody, fabrication, manipulation, advertising, and propaganda [Collins *et al.* 2020]. The overarching commonality extents to facticity and deception [Afroz *et al.* 2012].

Himdi *et al.* [2022] present a supervised machine learning model, providing the potential harm of fake news and how it became a critical concern in mainstream media. This model classifies Arabic news articles based on their context's credibility. It creates a unique strategy to form lexical word lists and develop an Arabic NLP device to perform textual feature extraction. From a marketing perspective, it is profitable to identify potential users or customers by characterizing them and understanding their reviews or comments on certain products [Weren *et al.* 2014].

Pennacchiotti and Popescu [2011] present an authorship attribution analysis to distinguish whether a Twitter user is a fan of Starbucks, which may apply to other businesses; however, most methods using online text snippets for business, politics, and social networks rely on considerable feature engineering to reflect the content and the author's style. These methods achieved reliable performance; however, they lack transparency and interpretation, essential to real applications, such as authorship identification, profiling, and plagiarism detection, requiring evidence or an explicit explanation.

1.1.3 *Literary science and education*

Authorship attribution is crucial and has had a significant impact in the digital humanities. Literary science and education are interested in determining who wrote historical literary works such as the Pickett letter during the Civil War letters [Holmes 2003; Ding *et al.* 2017], Shakespeare's playsciteplauriat2018literary,ding2017learning, the Federalist Papers citepmosteller1963inference, and the classic French authorship mystery of Le Romande Violette [Bogdanova and Lazaridou 2014; Ding *et al.* 2017]. The Federalist Papers are a collection of 85 political essays written between 1787 and 1788 by an unidentified author named "Publius" who pushed for the ratification of the United States Constitution. The authors of 73 papers were identified with consensus, but the remaining 12 were disputed using stylometric features rooted in statistical analysis to determine authorship based on linguistic patterns and writing style. As a result, Alexander Hamilton wrote 51 of the papers, James Madison wrote 14, and John Jay wrote 5; they were discovered as disputed authors. Additionally, there were four papers that could not be attributed to any specific author.

The same authorship attribution methodologies are equally effective for assessing the performance of literary translators and evaluating the accuracy of their translations. Translators will not have their own writing style but will replicate the original author's style [Ding *et al.* 2017]. Furthermore, the approach has been used to understand the personalities of kids based on their writing style in their native language and predict their future academic success [Almishari *et al.* 2014; Torney *et al.* 2012; Ding *et al.* 2017]. Authorship attribution approaches detect plagiarism and identify authentic authors of anonymous articles (Stamatatos, 2009; Juola, 2008). Furthermore, these methodologies can be used in a variety of sectors, including forensic linguistics and cybersecurity (Rocha, 2016). These measures have proven vital in protecting the integrity of written material and combating deception and fraud [Claxton 2005]. They have considerably improved the fields of forensic linguistics and cybersecurity, and their prospective applications are actively being explored. These technological breakthroughs have transformed the field of information security, holding great promise for the future of protecting sensitive data and avoiding cybercrime.

Mosteller and Wallace [1963] used combinations of conjunction words and prepositions, e. g."and", "of", and applied Bayesian statistical methods to measure common function words (FWs), which were used as a feature to determine the writing style and predict text authorship. Although the model achieved remarkable results, intrinsic meanings limited the chance of extrapolating the writing style from unidentified data that is part of the training set. The research study adopted the most common FWs to express abstract concepts as well as authorship, regardless of the topic. Similarly, Sarwar *et al.* [2018b] employs natural language processing techniques to find a pattern, e. g.a set of a few terms from the search queries to determine the

writing style from the document collections. However, the challenge of managing disputed texts remains, as multiple authors have similar writing styles or intentionally mimic another author's style. Therefore, the presence of anonymous or collaborative writers further complicates the task of accurately attributing authorship to a specific text. Additionally, Segarra *et al.* [2015] used function word adjacency networks (WANs) to find the author of disputed or questionable texts. The methods find the author's preferred writing styles using nodes as FWs and directed edges as the likelihood of finding the sink word in the order of the source word. The approach used the transition probabilities of a Markov chain to interpret and determine the actual authorship of the text by comparing the terms of their relative entropies. But this method repeats grammatical structures, like FWs, and makes noise in the writing style stronger, which will be added to word structures or patterns in ways that fail to be consistent with stylometric markers. This limitation leads to inaccurate results and hinders the effectiveness of the method in identifying the true author of a disputed text. Additionally, the reliance on Markov chains neglects other valuable linguistic characteristics that could contribute to an author's writing style, such as vocabulary choice or sentence structure.

1.2 PROBLEM STATEMENT

Texts are the only evidence available attributable to the incredible development of Internet media, despite all the challenges associated with authorship attribution. Internet media constantly contributes to the increased availability of user language and words in digital form. This study, therefore, informally describes the PAA challenge for online text snippets (OTS) and documents, followed by a formal problem definition. This study provides candidate authors of an anonymous, short (involving 280 characters [Isaac 2017; Gligorić *et al.* 2018] commonly established in social media posts). It includes long text samples (LTS) (containing thousands of words [Sarwar *et al.* 2018b], commonly established in a discussion forum and digital application [Keselj and Wang 2003; Ramnial *et al.* 2016a; Sarwar *et al.* 2018b]) with sample writings for each candidate author. Authorship attribution experts aim to analyses the anonymous text and the sample writing of each author. The analysis aims to capture the writing styles of the anonymous OTS or LTS and the sample writings of the suspected author.

As a result, the PAA expert must identify the most likely author of the anonymous OTS or LTS as a suspected author whose writing style is most similar to the writing style captured in the anonymous document. The goal of this study was to approach the PAA problem in an OTS or document, specifically e-mails, blogs, tweets, and online news articles that are perpetually produced from a variety of outlets on the Internet to deliver valuable content to people on a daily basis. Throughout the study, the terms "writing sample", "document", and "Tweet" are used interchangeably to refer to the same unit of text. Notwithstanding, post-authorship attribution (PAA), also known as post-authorship identification or categorization, assigns one or more labels from a preset set of labels or classes to a given document or piece of text based on the content [Deng *et al.* 2019a].

Formally, given a collection of N online text snippets (OTS) or long text documents (LTS) as $\mathcal{D} = \{d_1, \cdots, d_D\}$ and a set $\mathcal{T} = \{t_1, \cdots, t_k\}$ of k as a set of predefined author or categories of authors, a post-authorsjip attribution (PAA) problem can be constructed by mapping $\mathcal{F}$ as a function to model the Cartesian product of $\mathcal{D} \times \mathcal{T}$ to a set of binary $\{Yes, No\}$ or multiple class labels, e. g.$\mathcal{F} : \mathcal{D} \times \mathcal{T} \rightarrow$ {Yes, No} or multiple class labels, where $\mathcal{F}$ is known as a classifier. Based on this mapping, for a document $d_i \in \mathcal{D}$ and a class label or category $t_i \in \mathcal{T}$, if $\{(d_i, t_j) = Yes$, then d_i belongs to category t_j and need not belong to t_j or other multiple classes.

In modelling the pragmatically motivated feelings for author writing style in natural language for post-authorship attribution (PAA), which could be *subjective*[5], which is bias and inconsistency in the attribution process scientifically. Subjectivity can lead to varying results and a lack of objectivity in the assignment of class labels or categories. This lack of objectivity can undermine the reliability and validity of PAA methods, making it difficult to draw accurate conclusions about authorship.

To solve this problem, we classify a set $\mathcal{D}$ of OTS or LTS into various groups based on their content or traits and create a classifier $\mathcal{T}$ that can be used to figure out who wrote indistinguishable texts. Therefore, we train $\mathcal{T}$ by mapping it to a set of $\mathcal{D}^{'}$ documents. As a result, $\mathcal{T}$ does not know much about the cause-and-effect relationship between OTS or LTS content and the document class labels or categories. This means that each document in $\mathcal{D}^{'}$ that comes from the proposed methods in Chapter 4, Chapter 5 and Chapter 6 only relies on the patterns and features found in the texts themselves. The trained classifier $\mathcal{D}^{'}$ is then applied to classify $\mathcal{D}$ based on the patterns and features it has learned, after which each document in $\mathcal{D}$ is assigned or predicted to come from $\mathcal{T}$. The accuracy of the classification can be evaluated by comparing the predicted labels with the true labels of $\mathcal{T}$.

While doing so, some key research questions emerge, which are described further in Section 1.3. Finding the most appropriate method to differentiate authors independently from each other using the proposed methods that are capable of correctly assigning authorship based on these unique features found in the document instead of hazard judgement is part of the question. Another part is making sure that the proposed PAA methods work well on a large scale or are scalable. Furthermore, it is critical to examine how the contributing writing style retrieved from the OTS or LTS of the given document accused of the crime will be explained or visualised, and accuracy should not be the only criterion for authorship attribution.

1.3 RESEARCH CONTEXT

The research context in this section is critical for understanding the main challenges of existing features and methods that have achieved good performance, as well as the underlying reasons why these methods fail, and then using this to propose a deeper regularised neural network to quantify linguistic intuitions expressed in the text or document. This approach seeks to increase the accuracy and reliability of language analysis and authorship identification.

1.3.1 *Research challenge*

In this thesis, the vital part of the proposed methods is to determine the author of the anonymous text based on the learnable writing style unveiled from the text to improve the accuracy of PAA systems for forensics applications; however, for the general public to make wise selections, they should also know the capabilities of the suggested methods and technology. This is crucial because post-authorship attribution (PAA) methods for social media forensics are still underexplored, and active research is needed to push them forward. The PAA for social media forensics experienced these difficulties:

1. *Insufficiency of data for post-authorship attribution tasks* The literature pays little attention to data collection for PAA and other style-related tasks in order to find ways to continue research on authorship-based tasks. As a result of collecting data for PAA and other style-related tasks to find ways to expand research on authorship-based writing, we are unable to know how the writing styles of known authors differ from those of questionable writers

5 e. g.meaning that the assignment (writing characteristics) of a class label or category to a particular OTS or LTS may depend on the personal interpretation or point of view of the expert analyst

in the training sets. Previously, information for authorship attribution was gathered via asynchronous computer-mediated communication channels such as discussion chats, blogs, and newsgroups. Few studies have looked at data collection on synchronous mediums for PAA tasks like OTS. Because there is no gold standard benchmark dataset for PAA tasks in social media forensics, comparing the performance of different methods is difficult.

2. *Methodology to extract valuable writing styles* Each author has a different style and shares specific features with other writers from large datasets containing OTS or LTS documents of known and unknown authors. Even though the ability to write is profoundly linked to individual personality, interest, knowledge, and collaboration within friends, very little work has been done to take advantage of these relationships. As a result, optimising the author's writing style to capture the underlying semantic within the content of the given text and pinpoint the extent of the expressed element interdependence to understand the actual sequences of each author is unclear and ill-delineated in the previous authorship analysis tasks. This thesis demonstrates a mechanism to fill in that gap by looking at character-level sequences, subwords, and predefined word vectors. These are then mapped into a vector space using deep learning algorithms to classify and pull out personal traits from written text. These approaches aim to uncover patterns and correlations between linguistic choices, writing styles, and underlying personality traits, ultimately enabling a more accurate assessment of authors' characteristics. However, further research is needed to refine these methods and establish a comprehensive framework for understanding the intricate connection between writing quality and personal attributes.

3. *Learning attribution* involves addressing these challenges by incorporating advanced natural language processing techniques and machine learning algorithms. These methods aim to capture not only the surface-level features of the text but also the deeper linguistic patterns and nuances that are unique to each author. By leveraging large datasets and sophisticated models, researchers can strive to achieve more accurate authorship attribution results, ultimately enhancing the reliability of forensic investigations in legal proceedings. In literature, machine learning (ML) methods such as support vector machines (SVM), artificial neural networks (ANN), Bayesian models, and random forests (RF) were the most effective in determining who wrote a piece of text based on a simple statistic of ordered word combinations that was used to teach. This occurs because the approach can handle common text or words that exist in a dictionary or that a linguistic or forensic specialist is aware of [Baron 2016; Altakrori *et al.* 2018; Rocha *et al.* 2016; Alonso-Fernandez *et al.* 2021; Theophilo *et al.* 2021]. However, the approaches are time-consuming and should not be considered since authors may employ many different phrases or unconventional text, making it difficult to develop an appropriate vector representation of the author's writing style (or words). Many of the approaches lack the knowledge to understand and analyse the syntactic and semantic structure to distinguish single or multiple authors from different sources that the document was collected from, as well as appropriately lingusitic intuitions related to critical elements in text to classify the class or label belonging to the suspected text or documents, which could be used as evidence during a pre-trial forensic investigation in a court of law or before a jury.

1.3.2 *Research Questions*

The following are the relevant questions that this thesis has addressed:

1. *How* do we extract and analyze these features to gain a better understanding of the author's unique writing characteristics so that the classifier can access a text or document efficiently, and which OTS or LTS is comparable to suspected authors?

2. *Can* we determine whether a writing style is the result of multiple authors or a single author from the input text or documents like OTS or LTS? This process is crucial in determining the consistency and coherence of the writing style. It also helps in identifying potential issues such as plagiarism or unauthorized collaboration.

3. *How* can we explain and interpret the author's contributing writing style in a meaningful and coherent way so that a forensic investigator or ordinary person can understand and identify the potential author of the text or document accurately?

1.4 TECHNICAL CONTRIBUTIONS

In this thesis, we make the following research contributions:

1. To develop an exploratory statistical analysis that uses machine learning algorithms and visualization tools to examine the linguistic variation in the input text in a language-independent manner that can be useful to model inherent phenomena in author writing style for post-authorship attribution. This approach can help in identifying the unique linguistic patterns and features that are characteristic of each author's writing style. It can also aid in distinguishing between multiple authors who may have contributed to a single piece of text.

2. To investigate the use of character-level convolutional neural networks, bidirectional long short-term memory encoders, and distributed highway networks over a fully connected layer and softmax function for dealing with raw input text such as OTS and LTS in order to achieve satisfactory prediction on learning an individual's writing style. This method has a wide range of applications, including plagiarism detection, authorship attribution, and text classification. It can also provide insights into an author's distinct writing style and aid in differentiating it from others.

3. Developed an interactive system to visualize the contributing author's writing style in the post authorship attribution to assist in evidence-based forensic investigations. The system allows for easy interpretation of the data and can aid in identifying patterns and connections between different pieces of text. This technology has the potential to revolutionize the field of forensic linguistics and improve the accuracy of investigations.

4. In this thesis, the effectiveness of differently designed models is scrutinized on four different OTS and LTS datasets in English. A thorough experimental comparison using well-known methods is also shown. The outcomes indicate that the outlined methods are effective at identifying authorship and detecting deception, making them a valuable tool for law enforcement agencies. Further research and development of this technology may lead to even more advanced applications in forensic linguistics.

1.5 STUDY OUTLINE

This thesis is organised into parts with six chapters with the following descriptions.

PART I contains two chapters. Chapter 1 provides the introductory message of the thesis and the context. Chapter 2 presents the literature reviewed in chronological order.

PART II introduces an on exploratory statistical analysis to visualize sets of linguistic variation in order to uncover and detect interesting facts or phenomena about the data. By visualizing the data, we can recognize patterns, gain a better understanding of the context, and reveal the degree of association between one dependent variable and another one or more variables in the big or high-dimensional dataset discussed in Chapter 3.

PART III introduces PAA using regularised deep neural networks (RDNN) in one chapter. Chapter 4 presents an instance of a convolutional neural network, bidirectional long

short-term memory as encoder and decoder over a distributed highway network to learn the writing style representation from OTS to predict questionable text or document for PAA systems. The convolutional neural network was employed to extract lexical stylometric features entered into the bidirectional encoder to extract a syntactic feature vector representation. The feature vector was supplied as input to the distributed high networks for regularization to minimize the generalization network error. The regularized feature vector was passed to the bidirectional decoder to learn a writing style. A SoftMax function was used to predict the feature classification layer. The RDNN method was evaluated against SOTA methods, using four benchmark experimental datasets to validate its performance. Experimental results demonstrated the effectiveness of the method when compared to the existing trendsetting methods investigated.

PART IV introduces two bidirectional long short-term memory (BLSTM) with a 2D convolutional neural network (CNN) over a two-dimensional pooling operation to capture sequential writing styles for distinguishing various authors in Chapter 5. To determine the writing style representation, BLSTM was used to obtain the sequential relationship among characteristics with sub-word information; 2D CNN was adopted to understand the local syntactical position of the style from unlabeled input text. The study extensively evaluated the model that leveraged sub-word embedding; it was compared to state-of-the-art methods for an extensive range of authors. The study methods improved by 1.07%, and 0.96% on LTS and OTS datasets, respectively, and produced comparable results on the remaining state-of-the-art (SOTA) (ultramodern) methods.

PART V introduces an attention-based bidirectional long short-term memory and bidirectional gated recurrent unit with a convolutional neural network (ABC) model to circumvent the inherent challenges of PAA in . Chapter 6 presents a combination of long short-term and bidirectional gated recurrent units to extract a lexical stylometric from a distributed representation of words entered into an attention-based mechanism to emphasize various categories of syntactic feature vector representation. The feature vector was then supplied as input into a convolution and pooling mechanism to reduce the dimensionality of the features and learn the position-invariant of the author's writing style. A fully connected network and a SoftMax function are used to make the prediction. The ABC method was assessed against 20 trendsetting methods using four benchmark experimental datasets to validate its performance. Experimental results demonstrated the method's effectiveness compared to the trendsetting methods on two datasets while producing comparable results on the remaining two investigated.

PART VI presents the concluding aspects of this thesis and future directions for research in this area (see Chapter 7.

Part II

LITERATURE REVIEW

This chapter provides foundational information on authorship attribution and how it pertains to post-authorship attribution tasks. It includes a detailed description of the various features and methodologies used in previous literature as well as the experiments presented in subsequent chapters using the distinctively proposed approach. Furthermore, we present a variety of authorship attribution tasks, such as author profiling and verification. We also describe an evaluation forum that covers several authorship identification shared tasks and demonstrate how it aids in the advancement of authorship attribution research. The advantages of developing automated authorship attribution applications will help researchers and investigators quickly identify and eliminate cybercrime suspects. This chapter covers all aspects of authorship attribution research, from selecting features and preprocessing them to determining who wrote what. To address the aforementioned challenges, we investigate author classification based on online text snippets from social media such as email, blogs, forum posts, and chat logs, with a focus on high-quality training texts.

2 BACKGROUND AND RELATED WORK

In Chapter 2, we provide a comprehensive description of prior related work in the field of authorship attribution and highlight its significance in the context of forensic applications. We start by discussing the complexity of social media and the complicated task of identifying users of online text snippets (OTS) or long text samples (LTS) during a pre-trail investigation. In another section, we provide an overview of the basic characteristics that need to be extracted from a given text or document to understand and analyse the author's writing style, and various analysis techniques that have been used in the literature to attribute the individual to the related elements are explored, respectively. Finally, we outlined the main problems and challenges faced in this field of research.

To be coherent in nomenclature, in this Chapter 2, alternatively, we use the terms tweets, online text snippets, or documents as a union of available labelled or unlabeled posted messages as an online tesnippetest (OTS) or long text sample (LTS) from social media sites such as Twitter (or "**X**"), a blog, or the Internet. These posts take the form of a text dataset, images, or videos and are commonly referred to as a collection of the bare minimum unit of text data to be analysed for post-authorship attribution (PAA), which is required to identify the original author of a post or piece of content. The main purpose of these statistics is to enable accurate identification of the original author of a post or piece of content, which is crucial for various authorship attribution tasks. Furthermore, to refrain from mix-ups, we will use Twitter rather than "**X**", as it is formally known in the public domain.

2.1 SOCIAL MEDIA

Social media is a class of Internet-based apps that are built on the ideological and technological foundations of Web 2.0 and enable the creation and exchange of user-generated content [Humphreys 2017]. Social media has a direct impact on people's private, professional, and academic lives. There are numerous applications for social media, such as collaborative projects, microblogs, content communities, social networking sites, virtual gaming worlds, and virtual social worlds [Kaplan and Mazurek 2018]. Social media has some traits associated with interpersonal communication, yet it also shares other traits with mass communication. Social media usually, although not necessarily, involves the use of digital technology to facilitate some kind of public or semi-public communication. Social media is a popular platform for people to consume, share, and discuss information about scientific advances, politics, and controversies [Chen *et al.* 2022]. It is also a well-known phenomenon that the real lives of people on the platform can be difficult to determine merely because their online presence can appear to be a machine or a different person compared to how these individuals appear online [Rocha *et al.* 2016]. Moreover, nature and scope of social media make effective response and management challenging.

In a recent investigation and publication by New York Times Magazine[1], for example, there was a case of a Russian media agency that allegedly executed organised disinformation information campaigns on social media using pseudonyms to manipulate public opinion and spread false narratives [Rocha *et al.* 2016]. After a thorough investigation, it was proven that the agency had a substantial network of fake accounts and sophisticated tools to propagate

1 https://www.nytimes.com/2015/06/07/magazine/the-agency.html?auth=login-google1tap&login=google1tap

messaging to their intended victims through social media using pseudonyms and virtual identities. The investigation also suggests that the campaign was in fact funded by the state government's. With a staff of media professionals, the angency succeeded in disseminating false narratives, influencing public opinion on politics, and even deceiving the journalist covering the story for Time magazine.

On the Internet, these are trolling —- a collection of online behaviors that involve the use of antagonism, deception, and vigilantism, or any reconfiguration of these social behaviors to provoke reactions from people or institutions [Demsar *et al.* 2021]. It has become increasingly prevalent and hazardous on social media and online forums in recent years, posing significant difficulties for maintaining healthy and constructive online communities [Quandt *et al.* 2022]. Determining the people who wrote these online text snippets (OTS), reposting OTS, or even the original proxy address that the OTS were transmitted through could be important to prove the legitimate owner of the text, but these are concealed easily, making it difficult for a forensic investigator. But in all of these cases, one important factor stands out: the anonymity built into social media, which encourages people, including criminals, to carry out illegal activities online. This anonymity provides a shield for wrongdoers, making it challenging for forensic investigators to trace their actions back to them.

In a recent survey, over 25% of Americans[2] have engaged in social media trolling by sending nasty tweets to strangers [Gammon 2014], while one-third of Australians have been trolled online[3]. Many of the users concerned are tuning on the privacy setting on their account or employing a service, e. g.an SMS[4]-based one-time password [Mulliner *et al.* 2013] or routing services [Narula *et al.* 2008] that randomly check user data back and forth for anonymization. In the case of the Russian agent, the traffic was tunnelled through a freely available open internet proxy server[5] [Rocha *et al.* 2016]. However, gathering documentation from these can not be used as convincing evidence in a court of law, particularly when two or more experienced people are combined, and the aforementioned frustrates the investigator. In many of these cases, the written word e. g.text, is the only evidence accessible on social media to solve the quandary and the only clue to the author's identity. The text or data revealed in this particular incident could be about the alleged perpetrator's writing style and surrounding environment, which can be helpful in quantifying the statistical pattern in OTS or LTS as language that is linguistically modelled in natural language to understand and analyse the inherent characteristics. This helps to make the digital world a safer place for everyone while also protecting potential victims and ensuring that cybercrime perpetrators face justice. This will not only dissuade potential cybercriminals but will also foster a sense of security and trust in the online community. It will send a strong message that cybercrime is not tolerated. Furthermore, the constant evolution of technology and the emergence of new social media platforms only add to the complexity of uncovering the true identities behind these illicit activities.

2.1.1 *Twitter*

It is undeniable that Twitter, as a communication platform, has become increasingly ingrained in virtually everyone's daily life, irrespective of geographical location. Since the initial launch of Twitter by Jack Doresy on March 21, 2006, and when the very first message was posted

2 https://shorturl.at/ejzCO
3 https://www.sbs.com.au/news/article/one-in-three-australians-say-theyve-been-trolled-online/to8fjyyeq
4 Short Message/Messaging Service
5 https://incloak.com/proxy-list

on the platform, an estimated 3 trillion tweets have been sent[6]. Approximately two-thirds of these have been removed from the platform because the recipients deleted them or because the accounts (and all of their tweets) have been blocked from the platform or are otherwise inaccessible by means of the historic browsing facilities with Twitter's standard application programming interfaces (APIs), which are the most common entry points to collect data. Twitter has a number of publicly available APIs that enable free but limited access to data. There are two of them: the streaming API and the search API. Drawing on Twitter's official guide[7] and other research [Kim *et al.* 2020; Chen *et al.* 2022], the main features and differences are summarised between the two APIs.

The streaming API is constructed to retrieve tweets in real time. In fact, there are now a pair of options available: filtered and sampled streams. The filtered stream makes it possible for study participants to screen millions of new tweets per second and collect only those that correspond to a certain set of filter rules (for example, up to 400 key terms, 5,000 user identities [IDs], and 25 locations). By default, the filtered stream retrieves 1% of tweets per hour. If the study participant would like to use the filtered streaming API to retrieve all tweets with the hashtag #TwitterAPI, for example, and there are fewer matched tweets than the allowed cap, the researcher will get all tweets with that hashtag; otherwise, the researcher will get a sample of #TwitterAPI tweets [Chen *et al.* 2022]. In contrast, the collected stream is intended to produce a reasonable random selection of all freshly submitted tweets in real time, free of filtering constraints [Chen *et al.* 2022; Pfeffer *et al.* 2023]. The filtered stream is better suited for study participants who want to collect information on certain themes. When those conducting studies do not have a specific inquiry topic in mind and simply want to gauge the temperature of all Twitter discussions for general monitoring purposes, the sampled stream is a preferable tool to employ. Using Twitter's streaming API[8], we estimate that there were around 900 billion public tweets on the platform when Elon Musk bought it for $44 billion in June 2023[9].

The searching API is another popular entry point for accessing Twitter data. It was designed to return historical tweets collected by matching rules defined by users. The standard Search API enables researchers to access, at no cost, a sample of tweets published in the past seven days. For researchers who only need the past seven days of Twitter data or who are willing to collect tweets every seven days to trace historical data, the standard Search API may be a good choice. If the research project needs historical tweets posted over a longer period and/or beyond the free-tier sampling rate, one would need to use the paid premium plans of Twitter's Search API. Usually, there are two premium-level search API products that allow backtracking of tweets for a longer period, with more than 30 days tracing back to 2006, when Twitter was launched. Additionally, there are other user-friendly platforms that study participants can use to access data from social media, for example, Twitter (see Figure 1), but they are very expensive and provided by a third party due to the rapidly expanding social media market. Salesforce Social Studio, Crimson Hexagon, and Synthesio all have different pricing structures, data access types and caps, visualisation and interface, historic search availability, and data output.

Again, Twitter has been instrumental in studying human behaviour with social media data, and the entire field of computational social science and sociology (CSS) heavily relies on information collected from Twitter [Edelmann *et al.* 2020]. These include developing methods of detecting tension in online communities [Burnap *et al.* 2015], disaster management [Seddighi *et al.* 2020], and emotional analysis during the pandemic of COVID-19 [Mathur *et al.* 2020]. Over

6 We do not have an official source for this number; it merely represents an educated opinion from a collaboration of numerous researchers on Twitter.
7 For additional information, see https://developer.twitter.com/en/docs/twitter-api.
8 https://shorturl.at/hNP59
9 https://www.bbc.com/news/technology-63402338

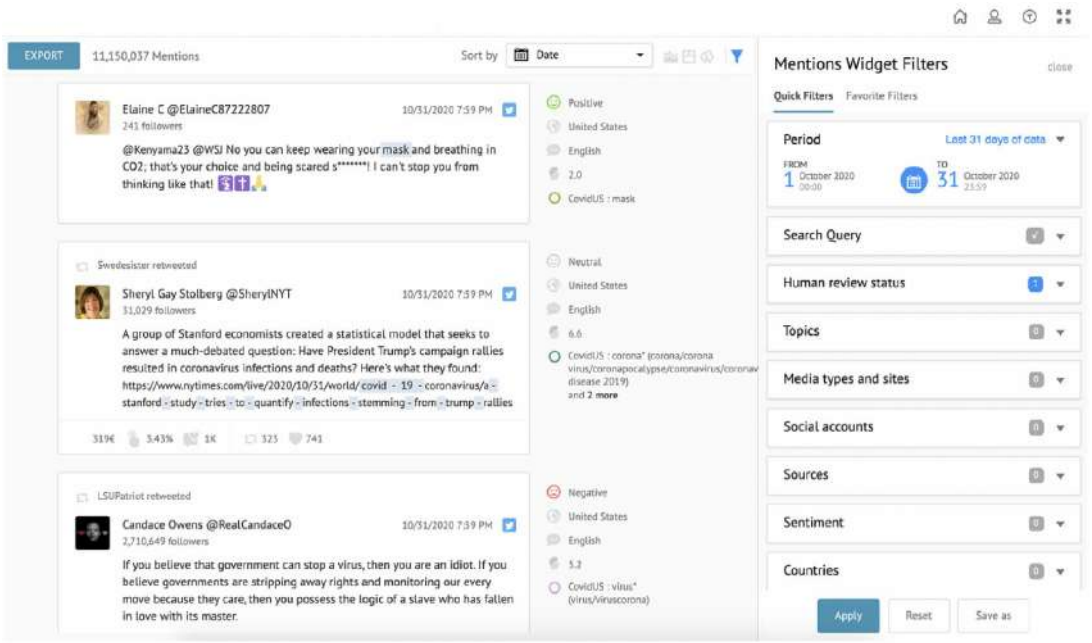


Figure 1: Layout platform of the third-party interface to access twitter [Chen *et al.* 2022].

30 scientific papers, including those presented at the AAAI International Conference on Web and Social Media (ICWSM) in 2021-2022, have been analysed on a subset of Twitter that include various topics, including public and mental health analysis, politics, and partisanship [Pfeffer *et al.* 2023]. It has further revolutionized the way people connect and share information, with over 486.0 million active users and over 400 million tweets posted per day [Kwak *et al.* 2010; Twitter, Inc 2022]. Twitter's global reach has grown significantly, and it is now available in 33 languages, greatly expanding its support for non-Latin character sets [Weller *et al.* 2014]. Twitter periodically modified the structure of the service [Oussalah *et al.* 2013], introduced new features, and discontinued the adoption of the API to crawl user-generated content (UGC) on the platform [Wang *et al.* 2019], either for research or commercial purposes. On the platform, users developed a conventional pattern of following subjective and non-subjective or trending topics that pertain to social, political, and all other aspects by adding '#' in front of important keywords to tag them, but the core concept remained the same: users are allowed to post concise texts (tweets) of between 140 and 280 characters in length, either based on follow-up posts uploaded by other users or a new post [Alshammari and Nielsen 2018; GligoriÄ‡ *et al.* 2018]. No doubt, Twitter data has enabled interdisciplinary research across many fields and has become a model creature for big data, but it has also been criticised for the various forms of discrimination that can emerge [Tufekci 2014]. Gaining access to data or the collection of posts on the platform to develop a preventive model rather than a reactionary one has been a significant disadvantage that has contributed to the issue of bias on the platform (now that the new owner has banned the crawling of information on the platform). This is a problem because data authenticity and reliability have never been guaranteed. While Twitter has long been one of the most collaborative social media platforms for data sharing with academic researchers, there is still a lack of transparency in sampling procedures, as well as potential biases caused by technical artefacts [Pfeffer *et al.* 2023].

2.1.2 *Characteristics of Twitter*

Twitter is a Web 2.0 tool[10] [Collis and Moonen 2008] that allows users or people to share or post meaningful conversations with pairs and share feelings as short digital texts, known as tweets. The user or people can post these tweets either directly or indirectly. Direct posts (personal posts or OTS) are used when a user wants to send an update to a specific individual, whereas indirect updates (public messages) are sent to anybody who wants to read them. Even though direct updates are designed to interact with a single individual, they are public and can be viewed by anyone. This feature of these tweets allows for both private and public communication, catering to different user preferences. Users have the flexibility to choose who they want to share their updates with, whether it's a personal message or a public announcement. It is important for users to be mindful of their privacy settings and choose their audience accordingly.

Most Twitter users, when posting messages or tweeting on the platform, keep a brief profile about themselves that includes their full name, the location, a short biography, and the number of tweets of the user. These tweets or messages can be tracked using phrases, words, and hashtags that are most often mentioned and posted regularly under the title of the trending subject matter or topics. A hashtag is a convention among Twitter users to create and follow a thread of discussion by prefixing a word with a '#' character. To collect user profiles, choose a user with the most followers and a large audience following them; however, Twitter limits the rate of crawling information or messages posted on the platform to 20,000 requests per hour. For example, we use a basic analysis to comprehend and visualise the number of followers and followings on the network based on the tweets collected in the thesis, as discussed in Table 3 and Section 3.2. Figure 2 shows the distribution of subscribers (solid line) and followers (dotted line). The y–axis depicts the corresponding cumulative distribution function (CCDF), and first, we show the distribution of the number of followings, which are noticeable glitches in the solid line. As shown in Figure 2, if it occurs at $x = 20$, the initial set of new users recommended by Twitter is 20, where followers follow the user by a single click and a large number of people take up the offer. Again, the second glitch is at around $x = 2000$ and this symbolizes the momentum of network building generated by the upper limit before it was removed by Twitter. A very enormous number of subscribers follow-more than 10,000. They are mostly the official pages of politicians and celebrities who need to offer some form of customer service. The dashed line in Figure 2, where $x = 10^5$, fits a power-law distribution with an exponent of 2.276. Most real networks, especially social media networks, have a power-law exponent between 2 and 3. The data points beyond $x = 10^5$ correspond to subscribers who have many more friends than the power law distribution predicts.

We went further to investigate the correlation between the number of followers and the number of written tweets in the Twitter dataset. Therefore, by evaluating the information using statistical methods and exploring the qualities necessary to identify influential users, we can gain insights into how information spreads and how trends in written tweets evolve on the platform. Figure 3, we plot the number of tweets (y) against the number of followers a user has (x), while we bin the number of followers in log-scale and plot the median per bin in the dashed line. The majority of users with fewer than ten followers never tweet or only tweet once; therefore, their count remains at one. The average number of tweets per user is always higher than the average number of followers, demonstrating that there are outliers who tweet significantly more than predicted based on their followers. The medium number of ntweets stays relatively flat in $x = 100$ to 1,000 and grows by an order of magnitude for $x > 5,000$.

10 The Web 2.0 innovation compass includes blogs, wikis, podcasts, tags, and social networks [Kasavana *et al.* 2010]

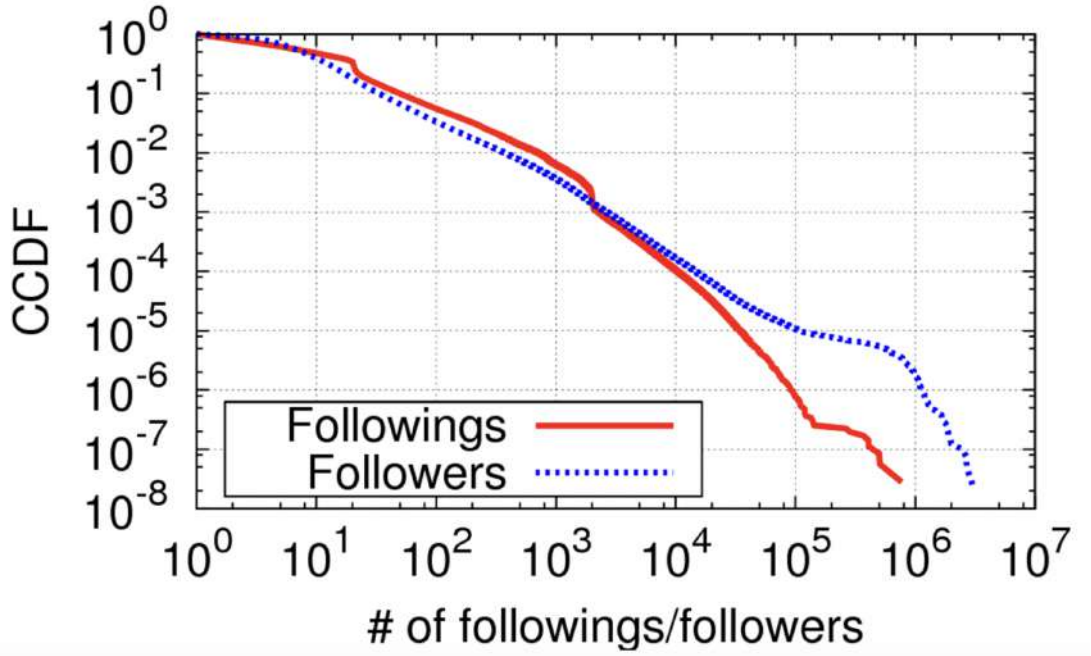


Figure 2: Number of followings and followers.

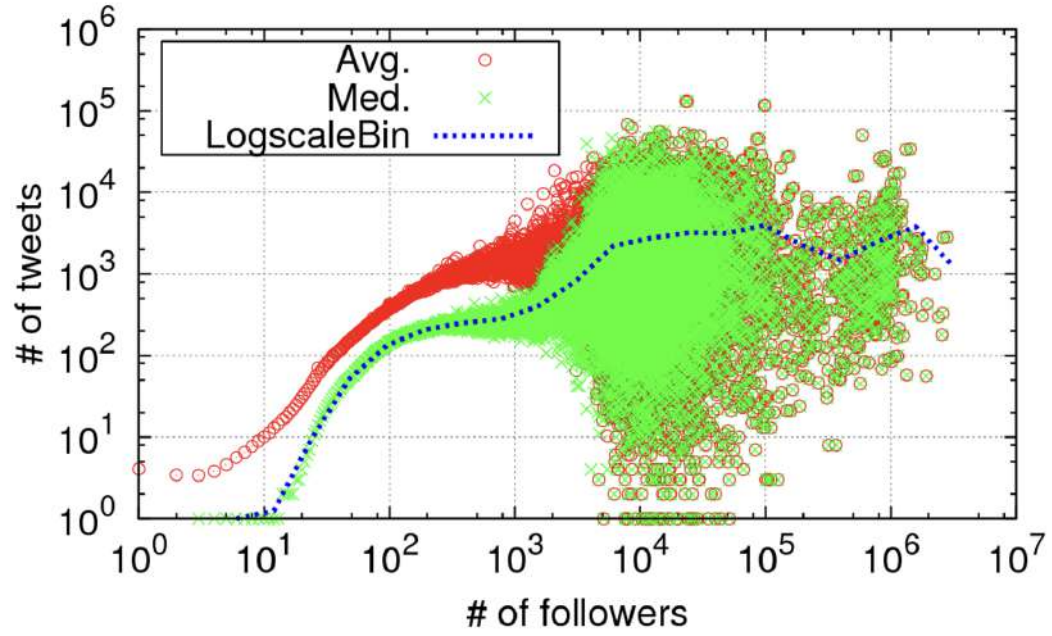


Figure 3: The total number of followers and tweets per user.

With open streaming or searching APIs available to craw, one can study the interactions and the retweet mechanism that users or people use in relaying information, which provides an unprecedented opportunity for computer scientists, sociologists, linguistics, and forensic investigators to understand and analyse human behavior elements in the tweets. Our proposed methods in this thesis contribute towards realising the full potential of determining the author's style of post-authorship attribution for an online text snippet (OTS) or long text sample (LTS) on social media. This research not only enhances our understanding of human behavior on social media platforms but also has practical applications in various fields such as cybersecurity, marketing, and public opinion analysis. By accurately identifying the author's style, we can detect fake accounts, track the spread of misinformation, and gain insights into user preferences

and trends. Furthermore, this knowledge can be used to personalize content recommendations and improve the user experience on social media platforms.

2.1.3 *Cybercriminal activity on social media via tweets*

Social media platforms on the internet are excellent for bringing individuals together to achieve a common purpose. Millions of people have used these online platforms as distinct communication avenues for real-time data sources that allow individuals to construct their own profiles and engage with other users regardless of geographical or physical boundaries. In this light, the online social network via online text snippet (OTS) has evolved into an indispensable communication tool. Communication over online social networks gives new insights into the creation of social networks and civilizations that were previously thought to be unattainable in terms of scale and scope. Moreover, social media network sites through OTS or LTS have transcended the boundaries of the physical world in studying human relationships and behaviors [Aggarwal *et al.* 2012; Al-Garadi *et al.* 2016]. However, a detailed examination of these relationships reveals the prevalence of unlawful events.

Cybercriminals have used social media platforms such as Twitter, Facebook, and others for many different endeavours, such as developing virtual communities, gathering information, and offering training and monetary rewards to members in order to propagate their illegitimate services citepresende2020framework. With the rise of online communication and social media, cyberbullying has become a significant issue [Al-Garadi *et al.* 2016]. Cybercrime is defined as the use of information and communication technology (ICT) by individuals or groups to attack other users [Salmivalli 2010]. Cybercrime is a substantial national health problem, with sufferers at high risk of suicide thoughts [Sampasa-Kanyinga *et al.* 2014;?]. Cybercrime, whether through cyberbuying or other parts of cybercrime, is a far more persistent version of old kinds of bullying that has a harmful impact on victims. The malicious attacker can harass his or her victims in front of an entire online community, and this has become an intrinsic part of the new world phenomenon. Therefore, social media platforms have become the most common places for cyberbullying victimization [Whittaker and Kowalski 2015], and their popularity and proliferation have increased the instances of cybercrime [Al-Garadi *et al.* 2016]. In general, social networking sites and the Internet or Web provide capabilities for cybercriminals to plan and execute their criminal acts. However, authorship analysis methods have proven to be the most effective strategy in recent years for identifying users, defending against such attacks, and addressing the problem of anonymity without compromising users' privacy[Altakrori *et al.* 2018; Majeed and Lee 2020].

Over the years, online text snippest (OTS) has been used or misused for a range of illicit activities, such as AFF [Modupe *et al.* 2011a], spamming, phishing, distributing malware, launching botnet commands [Zheng *et al.* 2006], and spreading malicious activities [Modupe *et al.* 2022]. For example, Figure 2 depicted a suspect involved in cyber espionage activities from an anonymous Twitter account dated 2000 with about 2,432 followers and 2,360 following. There was no clean posted message but a link to a suspicious website and OTS reading, "nAmAste Terrorist accord France An #CyberGuemillA. . . ". The anonymous user in this case creates social communities aligned with their goal of disseminating unlawful texts with images and adding malicious information references (a phishing website) to the tweets. Equally, the perpetrators are likely to collect information from social connectors on the page (e. g.users who follow and follow one another), and from the interactions, the account of the oblivious individual (e. g.the victim) could be hacked by the online criminals (such as when individuals re-post any of the tweets, reply to the message, or click the link). These are nefarious ways that cybercriminals hack into larger establishments to steal sensitive information or gain unauthorized access to systems. These actions create opportunities for online criminals to exploit vulnerabilities and carry out malicious activities.

Figure 4 shows an attempt at cybercriminal activities that create social communities to spread illicit texts and add malicious information references (a phishing website) to the tweets. Offenders may immediately collect changes from the social media pages of social connectors (e. g.users who follow and follow one another) and from the interactions of the hackers (such as when they re-post one another's negative tweets). This could make mischievous Twitter posts more visible and attract additional targets.

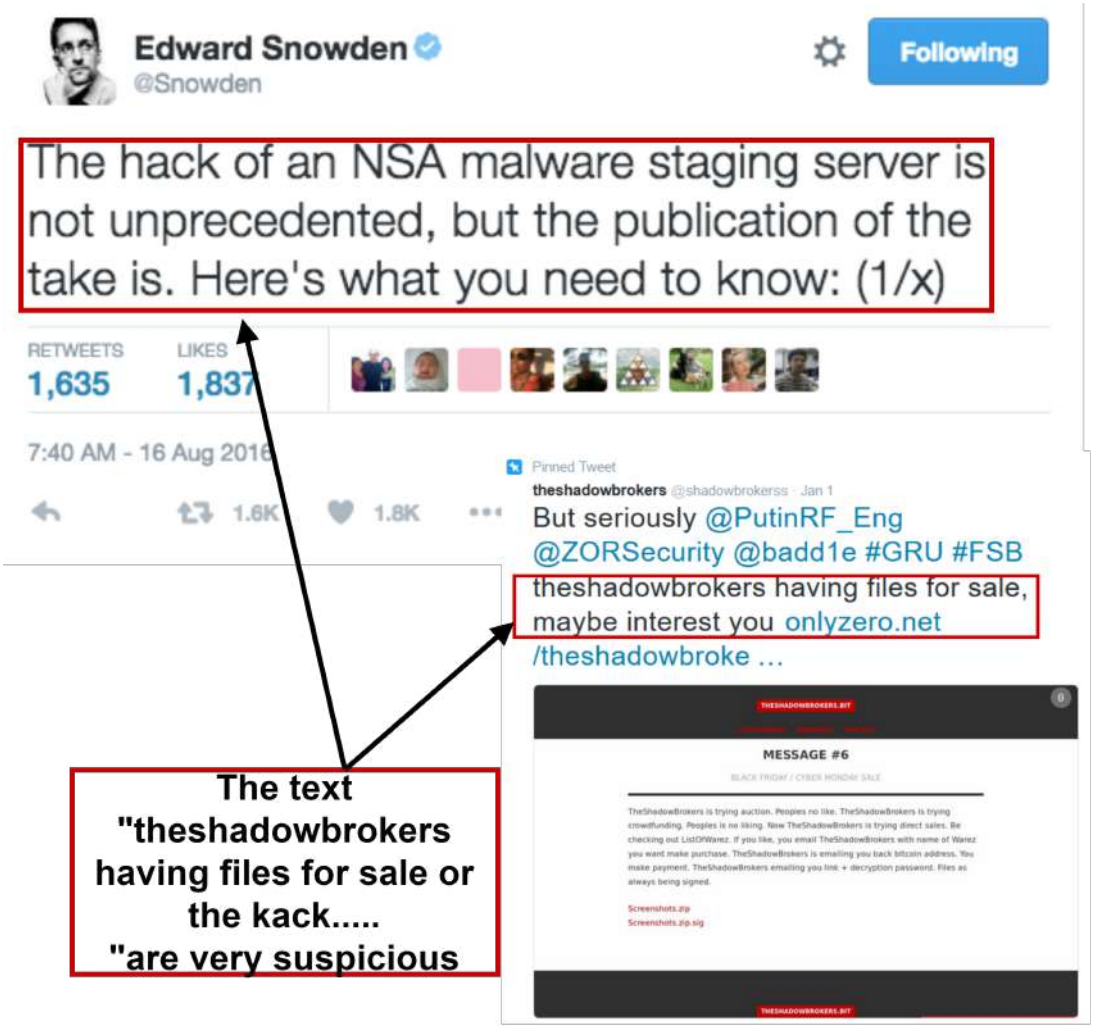


Figure 4: A online text snippets from hacker group.

2.2 FEATURE METRICS

In the course of an investigation prior to a trial, forensic linguistics assists lawyers in examining language-based evidence, such as texts sent anonymously to commit crimes online [Altakrori *et al.* 2018]. The PAA or AA task assists in resolving the issue of anonymity by determining who wrote the OTS or documents. The study of discovering hidden linguistic features in several text documents by known and unknown authors is known as authorship analysis [Venčkauskas *et al.* 2015]. It entails evaluating the writing styles or stylometric features of the text; however, in style-related tasks, data is collected by extracting and assembling parts of a text typical of that writer [Zheng and Jin 2023].

Characters are smaller units of words in texts, while multiple characters are amalgamated to form a commonly used collection of words [Valli and Lucas 2000; Khatun *et al.* 2019]. Words

can be grouped together in a variety of ways to form clauses, sentences, paragraphs, and entire texts. It is possible to determine who wrote an anonymous text by analyzing the sociolinguistic factors in the writing styles [Ding *et al.* 2017]. Semiotics, phonology, lexicology, part-of-speech (POS) tagging, syntactic, and rhetorical methods were chronologically ways to create unique fingerprints that could be used to facilitate the identity or quantify the characteristics of the author of a given text [Kalgutkar *et al.* 2019; Modupe *et al.* 2022; Zheng and Jin 2023]. Rudman [1997]; Stamatatos [2009b]; Koppel *et al.* [2009]; Neal *et al.* [2017a], over 1000 metrics could characterize the authorship of an anonymous online text snippet or document; however, no single metric is sufficient or universally accepted. These sections look at the main feature metrics used in modern AA or PAA tasks.

2.2.1 *Style*

Style measurements are measurable evaluations of a writing style that reflect the author's characteristics. Characters, word length, sentence length, and vocabulary richness are the various aesthetic indicators used in style metrics. The first attempts to determine the author of a questionable text, OTS, or document relied on simple measurement elements such as the length of each sentence, words, and hapax legomena [Mendenhall 1887a]. Hess *et al.* [1989] employed type-token ratios (TTRs) of oral language to measure the vocabulary-based diversity in each text, including duplicated words. These initial methods were limited in their ability to capture the complexity and nuances of a writer's style. As research progressed, more sophisticated techniques were developed to analyse writing style, taking into account factors such as syntactic patterns, semantic coherence, and even emotional tone. These advancements have led to more accurate and comprehensive evaluations of writing style, providing deeper insights into an author's unique characteristics.

Wu *et al.* [2021a] used 37 characters, 189 words, 12 punctuation symbols, and three sentence lengths to show different ways to simplify the linguistic metrics of a certain author from the text before looking at the ML model and figuring out who wrote it. In a similar manner Alonso-Fernandez *et al.* [2021] made use of 37 characters, 189 words as a total, 12 punctuation symbols, and three sentence lengths to portray various ways to simplify the linguistic metrics of an author from the text before exploring the ML model and attributing the authorship. The fact that style measurements (or style metrics) are not language-specific is a significant advantage. It can be used on any set of data, and the only thing that needs to be done is to make sure that the texts or documents that are being processed are tokenized (a device for turning text into tokens) [Stamatatos 2009b]. However, the large amount of text on social media sites made tokenization, the first step in processing text with NLP, more difficult. This challenge arises due to the informal nature of social media text, which often includes abbreviations, slang, and misspellings. Additionally, the brevity of social media posts makes it harder to accurately attribute authorship using linguistic metrics alone.

Some punctuation marks, especially the baseline dot, can be used in more than one way. For example, it can be (1) a period or full stop to mark the end of a sentence, (2) a marker for abbreviated forms, (3) a decimal mark to separate the integer from the fractional part of a number, (4) a separator between the hostname, subdomain, domain, or top-level domain in Internet addresses, or (5) part of a so-called horizon When these restrictions are anticipated, tokenization becomes more challenging as a task for humans [Proisl and Uhrig 2016; Zheng and Jin 2023]. The use of style measurements (or style metrics) to describe an author may be distorted or noisy if the vocabulary choice in the input text or sentence is misspelled [Neal *et al.* 2017a]. A single tokenizer, on the other hand, may work well on historical texts (a specific piece of text, such as a work of fiction, literature, or poetry, whose original purpose is to tell a story or amuse, as in a fantasy novel), but not so well on more contemporary ones, such as social media texts. As a result, it is best to select a tokenizer or dictionary based on the type of data and the purpose of the morphological analysis [Zheng and Jin 2023]. For example, a tokenizer designed for social media texts would be more effective in accurately identifying

and segmenting hashtags, emojis, and abbreviations commonly used in these types of texts. Additionally, a dictionary specifically curated for contemporary language would provide more accurate definitions and examples for modern slang and neologisms frequently found on social media. Therefore, understanding the specific characteristics of the data being analysed is crucial in choosing the appropriate tokenizer or dictionary for optimal morphological analysis. This ensures that the analysis of modern slang and neologisms is as accurate as possible.

Additionally, it is important to consider the context in which these hashtags, emojis, and abbreviations are used, as their meanings can vary depending on the conversation or platform. Incorporating contextual information into the analysis can enhance the accuracy of identifying and segmenting these elements. On top of that, regularly updating the tokenizer and vocabulary database to keep up with evolving language trends is essential for maintaining accuracy in analysing modern slang and neologisms on social media, but this will make the model or the proposed model not generalizable, increase the time consumption, and be inefficient for new features or elements that are unknown to the model in the future [Vosoughi *et al.* 2016].

2.2.1.1 *Word count*

A forensic investigator can use language-based evidence, such as the number of letters in a word or the length of the text or sentence in a particular dubious document, to identify the author. Mosteller and Wallace [1984], for example, used the presence of function words to identify the authorship of the contentious text in the Federalist Papers [Holmes 1994a; 1998; Stamatatos 2009b]. Likewise, Springer [1923] used four-letter words to identify the genuine author of Shakespeare's plays, whereas Bacon used three-letter words to formulate the facetious operettas under the counterfeit names of Shakespeare authors to protect himself from oppression. This approach of using specific linguistic features to determine authorship has been widely employed in the field of stylometry. For example, researchers have also utilised sentence length, punctuation usage, and vocabulary richness to distinguish between different authors [Argamon and Levitan 2005]. These methods have proven to be effective in various domains, including identifying the author of anonymous emails and detecting plagiarism in academic writing [Modupe *et al.* 2011b; 2012b].

By analysing patterns and frequencies of certain words or word lengths, researchers can uncover unique writing styles and potentially attribute texts to their respective authors. These methods have proven particularly useful in cases where the true authorship of a document is disputed or unknown. As a result of this observation, Williams [1975a] scrutinisedd the text of Philip Sidney, a Native English poet who distinguished his style of writing through his distinctive long texts against those established between Bacon's prose and Shakespeare's verse. When using word length to characterise the writing style of a particular author, it is critical to consider the impact of the linguistic form of some words in the text. For example, for some authors, the distribution of the word or length of the word in the text can be distinguished if assorted words depend substantially on the text's content; thus, to identify the author, it is critical to eliminate words that rely on the content. To identify the author, it is critical to eliminate words that rely on the content and focus on patterns that are independent of specific texts.

Word length and other statistical features have been widely employed to determine who wrote what and how they wrote it since the beginning of authorship attribution [Juola and others 2008]. Fucks [1952] demonstrated the complexities of word lengths in an attempt to determine different syllables among writers. Kelih *et al.* [2005] demonstrated the use of word length determinants in determining the originator of alphabetical characters based on the given text or who wrote the given text reached an accuracy of 70%. Grieve [2007] used word length distributions to distinguish two authors with an accuracy of 79%. Despite the fact that word length traits are widely used for PAA or AA tasks, they are primarily used to characterise a writer's traits rather than their writing style in order to improve PAA or style-based authorship-based tasks. The research community has harshly criticised the approach

as unreliable [Holmes and Tweedie 1995; Stamatatos 2009b]. But advancements in natural language processing (NLP) with deep learning (DL) and machine learning (ML) have shown promising results in accurately identifying the writing style of an author. These approaches analyse various linguistic features, such as syntax, vocabulary, and punctuation patterns, to capture the unique writing styles of individuals.

2.2.1.2 *Sentence size*

The number of characters or words in an online text snippet or document might influence the length or size of a sentence. Sherman [1888] used the length of a sentence in a text to exhibit a dissimilar average length to determine the true originator of a text. Yule [1939] employed fundamental statistical principles (mean, median, and word frequency) to calculate vocabulary richness. Like word length, there is insufficient information to make sentence length a viable method of achieving a high level of accuracy. Therefore, additional measures must be taken into consideration. One such measure is the analysis of the syntactic structure of sentences. By examining the arrangement of words and phrases syntactically within a sentence, it is possible to gain insights into the author's writing style and potentially identify patterns or unique characteristics that can help determine the originator of a text. Another approach is to analyse the semantic content of sentences, looking for specific vocabulary choices or thematic elements that may be indicative of a particular author or writing style. These additional measures provide a more thorough assessment and increase the accuracy of determining the true originator of a text.

Other stylometric methods have recently been introduced to increase the PAA system's performance. Sapkota *et al.* [2014], for instance, used an average number of sentences, tokens per sentence, punctuation marks per sentence, and 13 other stylometric features to identify the author of a document. Mekala *et al.* [2018] reported about 39 feature metrics, such as the number of sentences in a text, the number of words in a document, the average length of a word, the number of full stops, the number of exclamation marks, the number of commas, the number of colons, and the number of semicolons, to identify the author.

Wu *et al.* [2021a] presented a statistical style, indicating average words, sentence length, letter frequency, the number of 26 letters, and punctuation marks with three content features, two syntactic features, and one semantic feature to predict the author. Even though the length of a sentence is used to determine the author of an online text snippet or document, sentence lengths vary, complicating lawyer consultation through sentence length. This would serve as language-based evidence during a pre-trial investigation in court; therefore, sentence length is an inadequate descriptor to quantify the sociolinguistic characteristics of a particular instigator reflected in writing for social media text. Sentence length cannot accurately capture the sociolinguistic characteristics of a particular writer's online snippet text (OTS) or long text sample (LTS) on a social networking site. This is because sociolinguistic characteristics are complex and multifaceted, requiring a more comprehensive analysis.

Twitter, for example, encourages using abbreviations, short words, and slang in text posts online [Rocha *et al.* 2016; Savoy 2012; Naseem *et al.* 2021]. Furthermore, the main goal of networking is different from that of posting and lurking. People use social media to socialise and keep up with family members and acquaintances, whereas a major OTS, posting and lurking, is sharing information [Morrison *et al.* 2013]. On the other hand, criminals have used the Internet's anonymity to commit crimes covertly [Altakrori *et al.* 2018]. But since the start of Twitter, its abbreviation principle has been to engage, lurk, socialise, read, share, distribute, or talk about ideas and content in an open space [Geiger 2016]. Twitter has become a platform for both positive and negative interactions. This is because sociolinguistic characteristics are complex and multifaceted, requiring a more comprehensive analysis beyond just length measurement. In order to fully understand the impact of sociolinguistic characteristics on Twitter interactions, factors such as tone, context, and intent must be taken into consideration

and must be interpreted according to the specific cultural and linguistic backgrounds of the individual users involved.

2.2.1.3 *Vocabulary richness*

The vocabulary richness features of written material display how many words or phrases are used. They are unreliable because they depend heavily on the length of the text [Neal *et al.* 2017a; Shukla 2019]. Vocabulary richness with increased text length. Eder [2013] used the frequency of various words as a stylometric metric and de Vel *et al.* [2001a] used the number of tokens, the number of types (various words), the number of characters, and the number of hapax legomena to identify the originator of an email. Unfortunately, the vocabulary size depends on the length of the message (as the text becomes prominent, so does the vocabulary) [Stamatatos 2009b]. Baayen [2001], no metrics can bypass the influence of the sentence length; therefore, combined feature metrics could be employed to lessen the influence of text length and data structure on individual metrics [Zheng and Jin 2023]. Hoover [2003] presented a Burrow method considering large sets of high-frequency function words per 1000 words while applying principal component analysis (PCA) based on computing multiple lexical richness metrics as a standard method for AA tasks. Prasad *et al.* [2015] used an hapax legomena count of an average number of words; the number of sentences; the number of syllables per word; word length; sentence length; hapax dislegomena; enumeration of POS; tagging, and functional words to investigate authorship of Telugu text.

2.2.2 *Lexical*

Historically in literature, lexical metrics have been proven to improve the detection of sentiments and emotions in event detection tasks [Mohammad and Kiritchenko 2015; Thelwall *et al.* 2012]. It has been used to guide stylistic irregularities by quantifying the writers of a short or lengthy document [Ma *et al.* 2020]. Despite the clarity of the feature metric, it has the advantage of applying to any language once tokenisation of the text or document can be established [Stamatatos 2009b]. Lexical metrics regard vocabulary richness, word frequencies, word n-grams, and spelling errors. Vocabulary richness, representing the diversity of words used in a text, is unreliable because it heavily depends on the text's length. Vocabulary richness increases with the increase of proliferation of the size of text [Shukla 2019]. For example, Preotiuc-Pietro *et al.* [2019] introduces a systemic analysis of computational linguistics with the help of multi-perspective question-answer (MPQA) for lexicon sentiment [Wiebe *et al.* 2005] and the National Research Council (NRC) [Mohammad and Turney 2010] to the model expression of dissatisfaction of pain, or resentment across nine domains of Twitter users; however, the function of lexical metrics is limited as linguistic expressions in a long span of text often cause stylistic phenomena. According to Preotiuc-Pietro *et al.* [2019], the best-performing lexical metric in the experiments yields a worse result than a simple bag-of-words (BoWs) model; however, WordNet Oram [2001] is the most heavily used lexical database for finding semantic relation information between features and words. For example, Mao *et al.* [2018] uses hypernyms and synonyms of words from WordNet to disambiguate them when discovering nuanced metaphors from the text. Since writers express diverse attitudes towards specific events, lexical features can help solve simple cases in authorship-related tasks with high accuracy and, therefore, reduce 'noise' when training complex models.

2.2.3 *Syntactic*

Syntactic attributes can be believable if authors are inclined to use identifiable syntactic patterns instinctively and independently of the text for a given topic [Zheng and Jin 2023]. Getting syntactic features, on the other hand, depends on the language because there is a need for a

parser[11] to properly evaluate a certain intrinsic subjectivity taken from a natural language text. Noise is introduced as a result of unavoidable errors while extracting syntactic features [Koppel *et al.* 2011; Venčkauskas *et al.* 2015]. However, several studies showed that adding information about syntax could make the AA or PAA system more accurate [Baayen *et al.* 1996; Chaski 2005; Gamon 2004; Hirst and Feiguina 2007; Stamatatos *et al.* 2001].

Baayen *et al.* [1996] was the first to demonstrate the effectiveness of syntactic features by assembling more than a thousand rules to represent a portion of a syntactic feature in a text or document. The textual measurement of the rules was a feature. Short-text authors were distinguished using the bi-gram frequencies of a stream of syntactic labels [Hirst and Feiguina 2007]. These studies highlight the importance of incorporating syntactic information into authorship attribution systems. In summary, Baayen *et al.* [1996] developed a comprehensive set of rewrite rules that captured syntactic features in texts, and Hirst and Feiguina [2007] utilised bi-gram frequencies of syntactic labels to differentiate between authors of short texts Koppel and Schler [2003] presents an intriguing method by using syntactic misconceptions, such as sentence elements and incompatible tenses, as the attribution features. The most standard and straightforward methods are POS tag frequencies or POS tag *n*-gram frequencies [Sari 2018]. Even though some languages have similar styles, syntactic metrics show different features, like how words are separated or function words [Stamatatos 2009b; Venčkauskas *et al.* 2015]. This is a better way to do it than the old way of using punctuation frequency metrics [Hilton 1992; McMenamin 2002]. The goal of all of these methods is to find syntactic differences in texts that help identify the author's writing style from the given OTS or LTS. However, these differences alone are not enough to figure out what makes the author's writing style unique during a pre-trail forensic investigation [Venčkauskas *et al.* 2015].

2.2.3.1 *Function words*

Function words (FWs), which include prepositions, particles, determiners, and adverbs, serve as conjunctions for other words, indicate the grammatical structure, and have a relatively stable use. The FWs are designed to capture topic-independent functions and the pure intuitive stylistic decisions of the author in a text [Stamatatos 2009b; Zheng and Jin 2023]. Function words are important aspects of writing style and play a significant role in establishing coherence and cohesion in the text. FWs also contribute to the overall effectiveness and clarity of the written piece.

Golshaie [2019] found that FWs could accurately identify the writer's writing style with single-word phrases in diminutive sizes, and Hadjadj and Sayoud [2021] offered a contemporary collection of FWs using Arabic-language datasets by categorising 600 words into three categories, such as descriptive pronouns, authoritative pronouns, and conjunctions, in order to identify the authorship of the given text or documents by attaining an accuracy of 95%. When we look more closely at FWs and the methodological benefits to AA and PAA tasks, we can see three main uses: (1) FWs are reliable for textual comparison because they are defined in a group of phrases or words and are unable to be replaced; (2) their low dimensionality makes them useful from a quantitative point of view; and (3) authors do not choose FWs when they write. Instead, they are spontaneous and often emerge from the writer's subconscious mind [Kestemont 2014]. Therefore, FWs are used to determine who wrote a given text or document, which has become the primary indicator and focus of linguistic research. The ability to determine how these metrics are used for AA and PAA tasks without explanation or interpretation has been overlooked over the years, and this gap in understanding needs to be addressed in future studies. Nevertheless, the lack of conscious control makes FWs more creative and true to form

11 A parser is a syntactic analysis tool that analyses a string of symbols in natural language, computer languages, or data structures using formal grammar rules.

expressions of the writer's perceptions and emotions, which should be looked into [Kestemont 2014; Zheng and Jin 2023].

2.2.4 *N-gram*

Finding a *N*-gram of a string of a character, word, or token in a text or collection of documents is a common task in natural language processing (NLP) and one of the feature engineering techniques used in machine learning to extract features. *N*-grams are used to analyze the frequency and patterns of sequences of characters, words, or tokens, which can provide valuable insights into the structure and content of the text. They are particularly useful in tasks such as language modeling, information retrieval, and text classification. By examining *N*-grams, researchers and practitioners can gain a deeper understanding of the underlying patterns and relationships within the text data. Stylometric features for PAA or AA tasks, lexical textitn-grams, character textitn-grams, and POS tag textitn-grams are all textitn-gram metrics. It is critical to specify textitn when using the textitn-gram metrics, which can extrapolate the word count in a row or group of strings. Unigram metrics ignore word order (where $N = 1$). Character traits and linguistic N-grams, on the other hand, are better at capturing information about words, context, and themes; thus, the N-gram indicator is an excellent method to measure the author's writing style of a given text or document because some words or phrases are distinctive to how the author writes [Zheng and Jin 2023]. However, the weight of the *N*-gram metrics from a text or document could increase in dimension depending on the length of the text in the training data, which could exhibit exaggerated features. A couple of numeric values of *N* cannot represent contextual information in anonymous or questioned texts used for crimes. So, selecting the best *N* value depends on language, the task, and any among these methods [Neal *et al.* 2017a; Zheng and Jin 2023; Venčkauskas *et al.* 2015].

2.2.4.1 *Character N-gram*

Character *N*-grams are simple metrics for extracting smooth and continuous stylistic features in a text or document (e.g., various word forms) and are tolerant of 'noise' (or interference). Character *n*-grams metrics are because a text comprises a string of letters, numbers, capital letters, lowercase letters, or punctuation marks. Koppel *et al.* [2011], for example, used character *N*-grams to establish an appropriate author writing style that adequately characterized the author of a text or documents. Kjell [1994] used machine learning models (for example, neural networks and Bayesian models as classifiers) with the Federalist Papers by employing 36,000 character bigrams and trigrams to find the author's writing style and determine the individual author [Neal *et al.* 2017a]. This type of attribute is also available in any natural language model and is an excellent way to judge an individual's qualities of the model [Sari *et al.* 2017]. On the contrary, Sapkota *et al.* [2015] used character *N*-grams measurements to capture affixes and punctuation information, which has a significant impact on style-related tasks. At the same time, Barlas and Stamatatos [2020a] used character *N*-grams indicators associated with word attachments and linguistic punctuation marks, which are essential in determining an author across multiple contributors or subjects in a text or document.

One inspiring aspect of character *N*-grams metrics is how they use exclamation or punctuation marks. Howedi and Mohd [2014] extracted various lexical and character metrics of each writer's style using *N*-grams at the character and word level (where $N = 1$, 2, 3, and 4) as a feature to represent the input text. The study employed these features to examine the AA task on short historical Arabic texts written by ten individuals and achieved an accuracy of 93.33% only using character tri-gram metrics with punctuation mark features. Koppel and Schler [2003] captured common grammatical or orthographic mistakes (most common in an online text snippet), representing them as a character feature to identify the writer's writing style. The writer expressed that the text's effectiveness comes from using character n-grams to display content and style. This is comparable to what Sapkota *et al.* [2015] reported using

morphosyntax, thematic content, and style; however, commas remain a point of argument regarding punctuation [Wu *et al.* 2021a].

Commas are used not only in literary works, but they can also be used to identify the author based on the structure of the writing style in the text or document. The use of commas can reveal important insights into the author's thought process and help distinguish their style from others. Scholars and experts can gain a better understanding of the author's intentions and narrative techniques by closely examining the placement and frequency of commas. Additionally, commas can measure rhythm in a sentence as one of the metrics used in AA tasks; however, [Wu *et al.* 2021a] shown that the use of commas could appear several hundred times in given texts. As a result, figuring out the authors of OTS or documents such as diary entries is challenging, and it's extremely easy for non-writers to misuse commas.

2.2.4.2 *N-gram words*

Word *N*-grams are the most straightforward to understand. Word n-grams reveal the authors' preferred words or word combinations. Tan and Tsai [2010] explores the use of word *N*-gram metrics as particular characteristics to classify various authors for online text, including blogs and e-books. The findings show that syntactic features perform better with larger data sets, while lexical features perform better with smaller data sets. Howedi and Mohd [2014] found that word uni-grams outperformed word *N*-grams in terms of efficiency. The frequency of word *N*-grams proved to be more reliable and precise, according to Coyotl-Morales *et al.* [2006]; Sanderson and Guenter [2006] but the number of *N* continues to grow as the document or text increases. However, it is important to note that the performance of word *N*-grams may vary depending on the specific context or domain of the text. Additionally, researchers have also found that combining both syntactic and lexical features can lead to even better performance in certain cases.

The first drawback to extracting word *N*-grams from a text or document is that word segmentation, which is a grammatical phrase made up of one or more morphological elements or a means of determining the word boundaries in a text or document by computer algorithms. Therefore, for OTS or social media text, it is very difficult to extract or determine, not to mention developing machine learning algorithms to correct the error associated with word segmentation. Another obstacle with word *N*-grams is that unstructured data collected from OTS may be insignificant and irrelevant to the content of a given text or document based on the occurrence of words extracted from the text in order to determine the writer of the text or questionable text during pre-trail forensic investigation. This makes it difficult for a machine learning or classification model to accurately predict the author of language-based evidence regarding a writing style in court. Capturing content-specific information rather than stylistic information is possible; thus, determining the actual value of textitn is critical for content-based metrics or others composed of character and word textitn-grams. To investigate the authors' styles, highly dependent words must be removed from the content. Removing highly dependent words from the content allows for a more accurate analysis of the authors' styles. By focusing on content-specific information rather than stylistic elements, machine learning or classification models can better predict the author of language-based evidence in court. This approach enhances the reliability and effectiveness of content-based metrics, particularly those based on character and word N-grams.

2.2.4.3 *POS N-gram Tagging*

A computer-controlled grammatical interpolation of textual data is called POS *N*-gram tagging. It provides a potentially rich source of additional measurements that can represent an author's writing style for PAA or AA tasks [Rocha *et al.* 2016]. To do this, Gimpel *et al.* [2010] made a tag set, annotated data, developed characteristics, and reported tagging to trained English data from Twitter and achieved near-90% accuracy. It was compared to work by Altakrori *et al.*

[2018], investigating the authorship of Arabic discourse, particularly tiny text posted on Twitter. The study used three feature metrics, character, words, and POS, with an interactive device to visualize the similarity of the suspicious authors' profiles.Additionally, POS *N*-gram tagging is used to analyze and understand various aspects of the text, such as its complexity, linguistic patterns, syntactic structure, and semantic relationships. By examining these metrics, Wu *et al.* [2021a] aims to gain insights into the characteristics and properties of the text data being analyzed. Additionally, the metrics can also be employed to compare and contrast different texts or track changes in language usage over time. [Seroussi *et al.* 2010], for example, used it in conjunction with ML methods such as logistic regression to predict the authorship of a given text or document on the Internet movie database (IMDB) and achieved an accuracy of 87.74%.

Pokou *et al.* [2016] reported an adjustable dimension of POS *N*-gram tagging to generate author signature indicators, and Gómez-Adorno *et al.* [2018a] used the same features to get an accuracy of 91.9% when training datasets are encoded with POS *N*-grams by employing skip-grams (with the length of the input text ranging from 1 to 5). When compared to using word *N*-grams, the precision was relatively higher, with a 96.8% accuracy rate, but much lower when using character *N*-grams, with an average precision of 87.1%. [Litvinova *et al.* 2015] actually used POS *N*-gram to detect or identify author personality, and according to reports, a combination of POS and FWs were used as well. Citetjin (2014) showed the effectiveness of phrase parsing with POS *N*-gram including FWs, and punctuation to understand the surrounding content in non-English text. The benefit of these features was demonstrated by [Lee *et al.* 2017] using Korean text, but POS *N*-grams tagging is limited in the grammatical and contextual information it provides to genuinely determine the real author.

Dependency-parsing is combined with POS-tagging to help extract meaningful syntactic patterns. Basile *et al.* [2019] claim that pure POS tags are simplistic and can only model shallow syntactic features, whereas dependency parsing provides richer authorship information. Their experiments suggest that using POS and dependency features jointly accelerates proficient performance. The reason syntactic features are important in style-based tasks is that writing styles are usually specific to the choice and arrangement of words in a document; therefore, the interrelationship among words is indispensable. POS *N*-grams tagging indicators reveal insufficient sentence structure and contextual information, which is a two-edged sword. They reveal an author's distinct characteristics. In contrast, they only show the basic structure of sentences without describing higher-level structures such as phrases. However, the *N*-grams tagging criteria for POS distribution can vary depending on the textual input in the text or documents.

2.2.4.4 *N-gram phonetics*

To make phonetics reliable, there are certain rules about the cause-and-effect relationship between phonetics and pronunciations that have been used for AA tasks [Mosteller and Wallace 1963]. Deng and Allahverdyan [2016] used sixteen native English writers to show that a Dirichlet distribution can be used to measure rank-frequency connections for phonetics. This approach allowed for a more accurate representation of the relationship between phonetics and pronunciations. Additionally, the study found that incorporating contextual information further improved the reliability of phonetics in AA tasks. However, Dirichlet distribution was found to provide a good fit for modelling the distribution of phonetic units, but it was very difficult to find the correlations because they do not include the distribution of specific phonetics that could be used to figure out how much the author contributed to the text.

Khomytska *et al.* suggests that the choice of phonetic elements in a text can vary depending on the intended genre or style of writing but is achieved with 95% accuracy using statistical methods with stylistic features to measure the author's writing style. Additionally, these findings highlight the importance of considering the linguistic context and purpose of a text when analysing its phonetic features.

Zheng and Jin [2023] demonstrates how a language besides English text is examined for effectiveness by employing phonetic metrics to pin down authors by comparing them with already existing indicators. Support vector machines (SVM) and random forests (RF) were employed as the classifiers. With SVM as a classifier, they reached the highest precision of 96% with comma indicators, followed by phrase patterns with 95% precision, word-tag bigrams with 89%, and phonemes with 84%. By contrast, the accuracy results of RF with phrase patterns, the use of commas, word-tag bigrams, and phonemes are 96%, 95%, 89%, and 85%, respectively. However, phonemes appear to have challenges successfully capturing the writing style of Japanese or English writers. They may not accurately represent the nuances and subtleties of the language and may lead to inaccuracies in language processing tasks. Additionally, this can absolutely make it difficult for forensic investigators to accurately analyse and interpret text written in these languages. In order to improve the accuracy of language processing tasks and forensic analysis, it is crucial to explore alternative methods that can better capture the nuances and subtleties of Japanese and English writing styles. This could involve incorporating other linguistic features or utilising advanced machine learning techniques specifically tailored for these languages.

Phonetics, the study of sounds in human language, plays a crucial role in understanding pronunciation and speech patterns. By encompassing a wide range of sounds, phonetics allows for a more comprehensive analysis of linguistic diversity across different languages. Additionally, the relatively consistent number of phonetics in various languages suggests that these features are more stable and less prone to variation compared to other linguistic aspects like vocabulary or grammar. These ranges of phonetics in various languages range from 20 to 50 (for example, Native English has 44 phonetics, Chinese has 35 phonemes, and the standard number of phonetics in any European language is approximately 37), showing that these characteristics involve less distortion than other features such as lexical features. The stability of phonetics in different languages can be attributed to their physiological basis, as they are grounded in the physical production and perception of sounds. Additionally, phonetic features play a make-or-break role in interpersonal interactions as they contribute to the intelligibility and clarity of speech across different linguistic communities. The negative aspect is that phonetic grammatical structures and written languages change over time, so new phonetics may show up frequently or disappear, and the inherent subjectivity of the language in the text may mismatch [Zheng and Jin 2023].

As a result, phonetics are constantly changing, making it more difficult to learn and communicate languages. For example, forensics investigators or professionals that excel at AA tasks may struggle to adapt to new rhythm patterns or elements in order to communicate with people who speak languages other than their own. Furthermore, the subjectivity of language in written texts can lead to misunderstandings and misinterpretations, as readers may perceive the same words differently depending on their linguistic and cultural backgrounds. These problems underline the need for continued phonetics and linguistics research and training to promote language learning and communication across linguistic boundaries. Additionally, including cultural sensitivity and awareness in language interpretation can help to avoid misconceptions and encourage effective cross-cultural communication.

2.2.5 *Semantic metrics*

The PAA or AA task falls under the style classification category, an exciting subfield of text categorization focusing on understanding a writing style from the vantage point of the linguistics employed in the text or online text snippet as opposed to the content. In the literature, various feature metrics and classification methods were proposed, attempting to overlook the text while focusing on its style. Because of the limitations associated with this metric, modern applications proposed for advancing research on authorship-based tasks incorporate multiple metrics to cover more extensive information and achieve higher accuracy. Basile *et al.* [2019] presents practical linguistic analysis metrics, such as context-free production and

semantic relationship frequencies. This would significantly reduce errors over more commonly used shallow features, such as FW and POS trigrams for AA tasks. Using document instances with five sentences each, 10-fold cross-validation, and a frequency cut from five, the author achieved 87.63% classification accuracy using all features and 83.00% accuracy using FW and POS trigrams.

Stańczyk and Cyran [2007] used the feature selection method to categorize input texts as lexical and syntactic metrics as author referential integrity. However, the author achieved an average precision of 95.78% for AA tasks by using FWs with punctuation marks and ANNs to identify the author of a questionable text. This happens because lexical and syntactic metrics are used subconsciously to reveal an author's style. Abbasi *et al.* [2008] employed a mixture of stylistic indicators, such as descriptive words, idiosyncratic, the use of POS, phrases or word length, vocabulary affluence, and paragraph measurement, which could assist in recognizing individual authors based on their writing style, accomplishing an accuracy of 95.2% on eBay comments and 90.4% on Enron email. Gómez-Adorno *et al.* [2018a] employed document embeddings to understand characters, POS N tags, and word N-grams for characteristic types. Zhang *et al.* [2018b] used a multiple channeled self-attention network (MCSAN) to determine a number of feature indicators as a supplement to recognize authors for AA tasks and achieved an accuracy of 92.9%. Alonso-Fernandez *et al.* [2021] approached AA tasks for digital text and evaluated them with nine feature metrics categorized as lexical, structure, POS tag uni-grams, punctuation symbols, and syntactic to capture writing styles at various levels. The evaluation result with limited training and evaluating datasets on Twitter achieved an equal error rate (EER) between 20 and 25% with 93 authors from the information security and object technology (ISOT) dataset [Brocardo and Traore 2014]. When a high number of Tweets is employed, unigram accuracy reaches 97–99%.

Koppel *et al.* [2009]; Stamatatos [2009b]; Grieve [2007] highlighted a variety of indicators, including lexical and syntactical characteristics, that could be used to improve precision for AA tasks and eventually identify the author precisely. Others implemented these attributes to reduce complexity, such as avoiding overfitting on training data, reducing irrelevant features, selecting the best subset of features, and indirectly improving classification effectiveness. Although feature selection has always culminated in an "ad hoc mess", the feature set behavior was regarded as classifier-dependent and non-deterministic and the most significant problem in stylometry representation for AA or PAA tasks [Juola and others 2008]. Koppel *et al.* [2009]; Stamatatos [2009b]; Grieve [2007] highlighted a mixture of indicators to enhance the precision for AA tasks. Houvardas and Stamatatos [2006] adoption information gain and entropy (IGE), while Koppel *et al.* [2006a] employed the likelihood ratio to quantify the degree of available "synonymy" as a feature in a text, and Grieve [2007] use χ^2; however, the selection feature loses discriminating power; therefore, stylometry characteristics do not hold the same capabilities to distinguish semantic relatedness among features [Neal *et al.* 2017a; Sari *et al.* 2017; Sari 2018]. However, most of these models rely on selection features such as BoW measurement [Ding *et al.* 2017], which are unable to capture different modes of author writing style, not to mention different variations in the preferred choice of vocabulary to boost performance on AA or PAA tasks [Sari *et al.* 2017; Sari 2018]. As a result, a mix of feature types is typically opted for on a theoretical level to understand the author's writing style for a given textual online text.

2.2.6 *Application-specific*

Considering the application-specific feature metrics of texts in certain circumstances can be advantageous. For instance, exploring emails and assessing the structure, indentation, and signatures (e. g.layout) could provide valuable clues to an individual's authorial style. The benefit of comments, placing curly braces, and other non-core feature metrics help the analysis of source code (SC) in the same way. Canales *et al.* [2011] introduced combined stylometry and keystroke analysis for online quiz authentication. Brocardo *et al.* [2017] used application-specific to capture paragraph structure for short messages. The paragraph structure includes

the number of sentences per text block, the average number of characters, words, and sentences in a block of text, and the average number of sentences beginning with upper and lower-case; however, the author can change application-specific metrics, which may not work for all text parts. For example, Microsoft Word keeps track of the time spent on constructing a document, including the name and date and other essential information. This information is helpful for author analysis, such as determining when and what an author likes to write; however, it does not help a forensic investigator to understand the author's writing style.

2.2.7 *Multiple feature metrics combinations*

The aforementioned feature metrics have shortcomings, and future AA or PAA tasks should combine multiple metrics to cover all relevant information and achieve higher levels of precision. By integrating various metrics, AA or PAA tasks can ensure a more comprehensive evaluation of features and provide a more accurate understanding of their performance. This approach would address the limitations of individual metrics and enhance the overall precision of future assessments. Many researchers who thoroughly study language extensively use deeper features like context-free production frequencies, n-gram frequencies, syntactic productions, and semantic relationship frequencies to make much fewer mistakes than people who only use shallow features like FWs frequencies and POS trigrams [Gamon 2004]. It was found that support vector machines (SVM) and other machine learning methods make it easier to look at big feature vectors with lots of different feature sets, achieving a baseline accuracy of 45% using a literary text from Anne, Charlotte, and Emily Bront$\acute{}$e[12]. This helps get very good results in style-based tasks where classification is needed.

Juola *et al.* [2019] look at the characteristics of an individual's text, regardless of the language in the context of the author's text, and argue that cognitive and social traits are likely to persist. Therefore, the researcher demonstrates this using the tweets by the user and the language (both in English, Greek, and Spanish, believed to be written by a single author) by conducting two separate analyses of bilingual corpora using the same individuals and measuring the correlation using basic stylometric features e. g.Zipf curve, Hapax legomena, entropy, and many others. However, the most immediately apparent issue is removing replication and extensions. In the study, it is not clear that there is sufficient information in the features identified that is capable of determining the writing style of the actual author of a given text with a high precision that is enough to be useful for future investigation, so more work is needed in identifying and clarifying features.

Khalid and Srinivasan [2020] explores the linguistic style used to investigate language in online groups. The study used a narrow definition of style as well as a broad definition with 262 variables to analyse the linguistic style of nine online communities on three social media platforms that discussed politics, television, and travel. We discovered that communities do have distinct styles. However, the study found that style is a very good predictor of nine online communities of news aggregation websites with over one million subreddit memberships (F-score 0.952, accuracy 96.09%). On average, it is statistically equal to content-based predictions.

Pennebaker [2011] proposed that, due to the psycholinguistic nature of style, people with similar mental conditions will have similar language styles. Wang *et al.* [2017d] used Flickr data to identify people at risk for mental diseases using stylistic similarity. Zaman *et al.* [2019] examined Google search histories using LIWC traits to identify people with low self-esteem. Both employ style as an indicator of psychological status to provide a composite measure for interpreting the author's writing style. Hu *et al.* [2013] analysed the language styles of

12 https://www.theatlantic.com/magazine/archive/2016/06/the-brontes-secret/480726/

three broadly defined communities, such as Twitter, email, and blogs. These communities, however, do not represent individuals with shared interests, which is our focus. Hu *et al.* [2016b] demonstrated that people with similar occupations share a similar language style. Both use style as an indicator to determine psychological status as a combined metric for comprehending the author's writing style. However, individuals were selected based on their occupation and not on the basis of a collaborative feature or writing style in the 'community' structure.

Abbasi and Chen [2008] used a combination of content word words such as POS, word length, vocabulary richness, and sentence length to predict and identify the author of an online text, achieving an accuracy of 94%. Gómez-Adorno *et al.* [2018a] implemented document embeddings using accumulated knowledge on characters, POS tagging, and word n-grams to determine cross-topic and cross-genre writers to achieve the best results. Additionally, [Wu *et al.* 2021a] created a multichannel self-attention model to pull out multiple features for an author recognition system. This method proved successful with a precision of 92.9% of the time. It is important to keep in mind that combining different feature metrics does not automatically combine an abnormally large number of features. However, it does make the data less noisy, so it is very important to know which features to use and how to combine information from different feature types to get the accuracy you want for AA and PAA tasks.

2.2.8 *Feature metrics Summary*

In Section 2.2, we have been able to demonstrate a number of features in the literature for quantifying the linguistic intuitions related to author writing style in literay and online text for AA and advanced PAA. Furthermore, we addressed how important it is to have multi-level attributes that can be put together to strengthen broader ideas of style in a more structured and understandable way. This helps to comprehend how the author's writing style can influence psychological notions like formality, sentimentality, politeness, etc. by figuring out other possible causes, like topic. This model also tries to use various linguistic features, such as character-level sequence, subwords, and pre-trained embedding vectors, to uncover unique pieces of author writing style altered without affecting the order of the text. By analyzing these features, we seek to accurately differentiate between different writing styles and determine if a particular style belongs to a suspected cybercriminal during pre-trial investigations. Additionally, this thesis examines the efficacy of deep regularisation strategies in enhancing and developing the neural network model that has the ability to generalise across various writing styles and improve its overall performance in identifying potential cybercriminals. Table 2 provides a chronological breakdown of feature metrics.

2.3 AUTHORSHIP ANALYSIS

Authorship analysis is the process of examining the characteristic features of a piece of text in order to conclude its authorship, which has its roots in linguistic research called stylometry [Koppel *et al.* 2009; Stamatatos 2009b]. Different authors have distinctive styles or means of writing as well as speaking, and there exists a long history of linguistic and stylistic investigations into authorship attribution [Swain *et al.* 2017]. Because of this, authorship analysis and attribution have different economic effects and are being looked into as possible areas of research in the field of stylometry. Stylometry is the study of finding the author(s) of any given text by using different text data mining approaches or techniques. Obviously, the long sample text (LTS) has better identification accuracy, which is anticipated due to the structured manner in the subject area of stylometry. However, authorship analysis has several other categories in the field. These include (i) authorship attribution (AA) or post-authorship attribution (PAA), which is the process of figuring out who wrote a group of texts; (ii) authorship verification (AV), which is the process of figuring out if a text or group of texts was written by a certain author; (iii) writer profiling (WP), which is the process of figuring out who wrote a text without

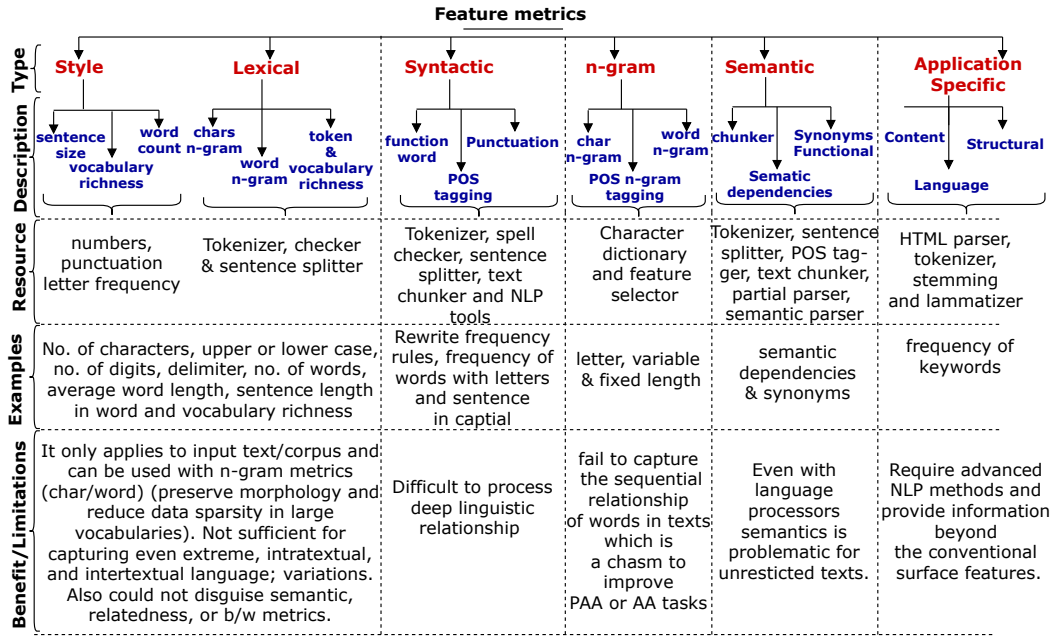


Table 2: Basic features metric for authorship analysis [Neal *et al.* 2017a; Sari 2018].

the author of the text; and (iv) style change detection (SCD), identifying text positioning within a multi-author document at which the author changes. Figure 5 depicted the aftermentioned task and post-authorship attribution (PAA), a subset of authorship attribution (AA), which is the primary focus of this thesis, and summarised the others as follows:

- The first problem is *closed-set* authorship attribution (AA) or post-authorship attribution (PAA). The purpose is to determine the genuine writer of OTS and documents, such as blogs, Tweets, and news [Modupe *et al.* 2022] based on the extraction of stylometric features demonstrated in **??**. The case is a multi-class classification problem where the machine learning model figures out and determines the right class for a text based on the learnable writing style (see Section 2.3.1).
- The second problem involves author verification (AV), also known as similarity detection and an open set of AA tasks. AV is a two-way problem that decides if the same person wrote two documents. The challenge is to verify whether a suspect is the creator of a text (Section 2.3.2) [Halvani *et al.* 2016].
- Authorship characterisation, which involves author profiling (AP), is the third problem. In this situation, a machine learning model or classifier must identify the compiler of a text among hundreds of potential writers, using the information on a potential suspect. The task is challenging because the classifier receives inadequate training data for each writer (see Section 2.3.3).
- Style change detection problem is to find if the writing style of the document has changed. The goal is to investigate whether it is created by one or multiple writers (see Section 2.3.4).

2.3.1 *Authorship attribution*

Authorship attribution helps to determine the writing style of the individual who wrote a text or digital content. However, stylometric methods in natural language processing (NLP) have undergone thorough investigation combined with many features outlined in Table 2. For example, Kjell [1994] was the first to use n-tuples for n from 1 to 5. The relative frequencies

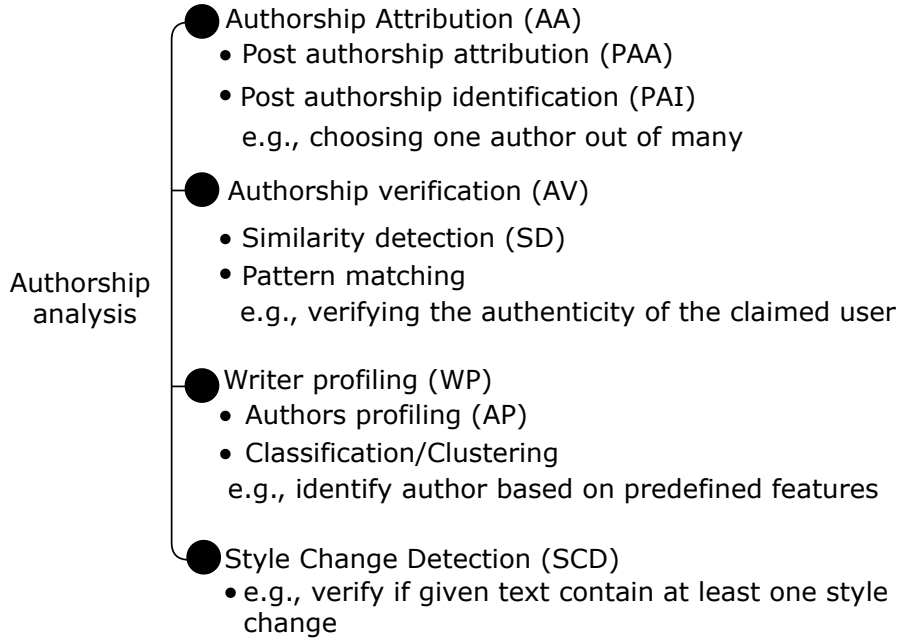


Figure 5: Authorship analysis tasks.

of the tuples were transformed with the KL-divigent and used as feature metrics. Then, the author fed the feature with a three-layer neural network and Bayesian classifiers to differentiate between text samples from two comparable writers of Federalist papers. However, the most effective combination was 2-tuples plus a neural network classifier, although other combinations performed similarly. Forsyth and Holmes [1996] showed that bigrams and character *n*-grams of different lengths worked better than lexical parts for classifying text, figuring out who wrote it, and extracting stylistic features. However, the number of retrieved feature metrics was limited to a shallow level of 96 features. All the same time, Keselj and Wang; Peng *et al.* [2003a]; Kešelj *et al.* [2003]; Stamatatos and others [2006] reported exceptional performance precision using character n-gram metrics of variable length frequency. Additionally, the character n-gram representation was one of the best-performing feature metrics in an ad hoc authorship attribution competition[13] in 2004 [Juola *et al.* 2006]. Similarly, Grieve [2007] compared lexical and character features on the same evaluation data sets and discovered that character n-grams were the best way to measure. In those experiments, only a mix of common words and punctuation marks did better.

As *n* is important in the definition of character *n*-gram methods, defining how long the strings should be is an important consideration for the success of the method [Stamatatos 2009b]. For example, a large *n* would better capture lexical and contextual information while also gathering more theme information. A small *n* (i.e., 2 or 3) would be able to express subword (syllable-like) information but not contextual information [Stamatatos 2009b]. However, [Forsyth and Holmes 1996; Houvardas and Stamatatos 2006] maintained that the extraction of n-grams of variable length can avoid the issue of defining a fixed value for *n*. While, Sanderson and Guenter [2006], the use of sequence kernels based on variable-length character n-grams yielded the best results for short English texts when examining sequences of up to 4 grammes. Several different Markov models with changing orders can be used to deal with character-level data

13 https://dh-abstracts.library.virginia.edu/works/365

[Khmelev and Teahan 2003; Marton *et al.* 2005; Stamatatos 2009b]. In conclusion, Zheng *et al.* [2006] constructed a suffix tree representing all possible character *n*-grams of variable length and then extracted groups of character *n*-grams as features.

With well-known techniques like machine learning [Juola and others 2008; Jockers and Witten 2010; Koppel *et al.* 2009;?] to look at different traits in social media text or online text snippets (OTS) using the feature metrics in **??**. The study tries to learn more about things like sentiment, phylogenesis, authorship, and authenticity analysis [Theophilo *et al.* 2021]. Nevertheless, the problem of OTS from social media is still not solved enough to be used in the real world because it is not as accurate as the state of the art (SOTA). For instance, Sapkota *et al.* [2015] found that affix and punctuation traits provide the best discriminative information for authorship attribution. The author analysed texts from news sources using support vector machines (SVM) and Naive Bayes (NB), with a maximum number of authors (e. g. 50).Rocha *et al.* [2016] presented an exhaustive bibliography survey of the SOTA techniques for PAA tasks. They employed shallow classifiers, such as SVM and RF, allied with a BoW feature representation. They explored the performance dependency of their proposed method with multiple writers and messages; however, their results are still far from satisfactory, despite advancing the SOTA for PAA tasks. The authors presented a classification accuracy of less than 65% in a new experimental dataset of 50 authors. This number plummets to less than 45% in 1000 authors, indicating that PAA is still an open research endeavour that requires further improvement. Most of these methods rely on discrete feature representation, which suffers from data sparsity and disregards the underlying semantic connection between the feature vectors for AA tasks. Alonso-Fernandez *et al.* [2021] approached the AA of digital text using stylometric metrics in **??**, among the widely used features in traditional literary analysis and more modern ways of writing on social media. They capture properties of writing style on various levels, such as the number of individual characters, the frequency of individual characters, the frequency of the sequence of characters, and the structural and syntactic metric of the text or online text snippet. They achieved an accuracy of 97—99% using unigrams or *n*-grams for tweets ranging between 1000 and 20.

Sari *et al.* [2018a] investigates how diverse feature metrics could affect the accuracy of the AA model under varying conditions. The author used style metrics, such as FWs, letter frequency, and punctuation, to determine a document's creator. The author used bags of *n*-grams combined with character *n*-grams to find the author's topical preferences on widely used datasets. The methods ignore the rich features that characterise the user's writing styles in the text. These features only reveal statistical information about documents in style, lexical, and syntactic groups. The limitation of the metric is that the feature-extracting process ignores rich sequential information and word order in the text. Again, Escalante *et al.* [2011] look at the word order in a short context using BOW. But the approach does not work for AA, PAA, or text classification tasks because the data is too sparse and has too many dimensions [Liu *et al.* 2020].

In the past few years, ANN could extract stylistic features from text and other objects [Ferracane *et al.* 2017; Hitschler *et al.* 2017; Shrestha *et al.* 2017]. In this method, the sequence of characters and words is the input of a neural network, making the proposed models use sequential information. Ge *et al.* [2016] employed a feed-forward neural network (FFNN) model on an AA task. The work reaches promising results compared to the traditional language model of *n*-gram baseline. Ruder *et al.* [2016a] proposes a hybrid CNN method that leverages character and word sequences for the large-scale AA task. The proposed method represents characters as one-hot vectors (discrete representations). Shrestha *et al.* [2017] employed CNN based on the character *n*-gram to determine tweet writers. The process takes a sequence of character *n*-grams as input. This model captures local interactions at the character level, unlike the last neural network models, using either a sequence of words or characters as inputs. Some years ago, Liu *et al.* [2020] talked about how to use a two-dimensional encoder representation that has already been trained from the transformer architecture with a dense layer and a SoftMax

function mechanism for AA tasks. The model's accuracy went up by 5.3% compared to current SOTA methods. To help with the author recognition task, Wu *et al.* [2021a] experimented with a multichannel self-attention network. This network uses interactions between channels and between positions to find personalities in n-grams, word n-grams, phrase path n-grams, topics, and dependency n-grams (where n = 1, 2, 3, and 4). They got a 92.89% accuracy rate with logistic regression on a subset of the Reuter Corpus Volume 1 (RCV1) datasets [Stamatatos and Koppel 2011]. Even though the results are promising, it is impossible to visualise or interpret the features and how they contribute to pinpointing the authorship of the text [Zheng and Jin 2023; Shrestha *et al.* 2017].

This thesis begins by discussing how character-level internet text samples can be utilised to learn the proper method to write like the author. To do this, a convolutional neural network, a bidirectional long-short-term memory encoder, and a distributed highway network that can quickly gather both global and local textual semantic vectors are combined. In summary, we are able to use the proposed model to deduce the abovementioned features of the author's writing style and visualise the contributing elements, which were discovered as research gaps in the literature, and hence, assist forensic investigators during pre-trail in the Court of Law.

2.3.2 *Authorship verification*

Authorship attribution attempts to determine the writer liable for an unlabelled text despite several classified documents, whereas verification is mostly a binary class concern by determining if the same writer created the presented documents [Halvani *et al.* 2016]. Authorship verification may not provide better information about the writer's attribute in a document but more about the variance between a pair of unidentifiable textual texts; however, some studies considered authorship verification by comparing the document's stylistic features.

Koppel and Schler [2004] indicated that authorship verification is challenging because a candidate author knowingly or unknowingly can arbitrarily shift writing style on the account they want to mimic an influential personality. The writer is conducting this for a personal reason or fame; it is challenging for the authorship system to determine the actual writer. Studies are restricted in authorship verification; however, published works share the approaches with those applied for plagiarism detection [Stein *et al.* 2007; Stamatatos 2009a; Zechner *et al.* 2009]. Most plagiarism software, such as Turnitin, Viper, Plagium, PlagTracker, and PaperRater [Naik *et al.* 2015; Ramnial *et al.* 2016a], used no form of stylometric feature to investigate copyright in the document. Most approaches cannot be used for AA tasks except to trace similarities in the document.

Stamatatos *et al.* [2014], authorship verification may be intrinsic or extrinsic. Intrinsic methods only operate on the input texts (known and unknown individual texts) to ascertain if the same author composes the document. Most published studies fall into this category that employed ML and similarity detection methods. For example, regarding a pair of documents—the similarity method would attribute the pair to a candidate writer if the similarity score surpassed a particular threshold. The scores can be determined using cosine similarity, Euclidean distance, min-max measure, and Manhattan distance. The approach is simple, though unsuitable for the AA task because style, such as topics, genre, and text shifts, are disregarded. The approach calculates the similarity score between the two documents. Conversely, studies were conducted with the intrinsic approach employing a supervised ML classifier with pairs of a classified document to create models; the pairs were then combined to distinguish the unlabelled text.

Fréry *et al.* [2014] used this approach by implementing a machine learning method that focused on multiple text representations and optimised with decision trees on diverse stylometric features and trained the corpus independently for the AA task. Their result was adjoint a 2nd position for the PAN CLEF competition in 2014 with an area under the ROC of 70.7% compared to [Peñas and Rodrigo 2011] of 68.40%. ML approaches for authorship verification

include SVM [Koppel and Schler 2004], clustering algorithm [Jankowska *et al.* 2013], and fuzzy **C**-means [Modaresi and Gross 2014].

Compared to the intrinsic technique, extrinsic methods direct authorship verification as a binary classification task. They attempt to generate a wide range of training samples as an impersonator or distractor text(e.g., a negative sample) [Stamatatos *et al.* 2014; Koppel and Winter 2014; Sari 2018]. For example, a pair of documents would be identified as written by a similar writer if the similarity score between the score from the documents is more significant than that of the impostors [Seidman 2013]. Several authorship verification methods employed the ideas of extrinsic techniques [Khonji and Iraqi 2014; Mayor *et al.* 2014] and achieved better results than the intrinsic methods. Koppel and Winter [2014] investigated the effect of three writer verification methods, including the similarity-based, supervised and impostor methods. The results indicate that the impostor system outperformed two other systems with an accuracy of 87.40%. The limitation of the impostor approach is how to determine the cross-validation methods to evaluate the datasets.

Koppel and Winter [2014] indicated that the quality of the impostor, the size, and the threshold score need to be optimised to obtain the proper equilibrium among the dataset during training. Further, the impostor approach underperforms when there is a slight change concerning the writing style in the documents. This study explored charCNN to detect character n-grams sequence over the input document by various filters and capture any change in sequence. It is anticipated that each filter shared across the receptive field over the convolution operation learns to activate various morphemes and construct a stylometric representation vector of writing styles from the identified word sequence. Although the representation perceived from the charCNN layer would conflict with the valid writing style in the document and be sensitive, the output was trained over the time distributed and BLSTM to capture contextual hidden linguistic styles for each sentence in the documents. The approach was extended using an on-line text snippets (OTS) to learn variation in the author writing style. The writer and precise feature influencing the change, or the writer was analysed.

2.3.3 *Writer profiling*

The writer's profile approaches demographic characteristics using several features, including native language, employment, gender, and age [Huang *et al.* 2014b; Ramnial *et al.* 2016b; Reddy *et al.* 2016]. Authorship profiling can detect emotions in communication, how they relate to age and apply to learn if the author is an adult or teenager [Peersman *et al.* 2011; Ding *et al.* 2017]. Profiling a writer is a daunting task for social media text as the features to classify the candidate can sometimes be unreliable, anonymous, or deceptive; however, some works focused on using indigenous language to interpret the writers' habits.

Koppel *et al.* [2002] employed a significant number of content-independent characteristics comprising 405 FWs and n-grams POS to differentiate two genders in British national corpus (BNC) [14]. They reached 82.6% accuracy when combining FWs and POS n-grams with prose-writing documents. Johannsen *et al.* [2015] conducted a large-scale experiment by finding a diverse syntax feature across several languages with various demographic statistics. The results indicate that men employ numbers and nouns often in their writing, whereas women use conjunction verbs more regularly. Experience revealed that younger generations use nouns frequently, whereas older generations use the prepositional phrase more often. Regarding age, younger groups use noun language more often, compared to adult age groups, using more prepositional expressions in their writing.

14 https://www.english-corpora.org/bnc/

Schler *et al.* [2006a] examined the significance of a writing style and the effect concerning age and gender among writers by incorporating close to a million words using user feedback on a blogging site. They implemented content based on a style-related feature, such as POS, FWs, and content-specifics selected from the LIWC dictionary [Pennebaker *et al.* 2001] to learn the writer's personality, age, and gender, based on vocabulary usage. They discovered that male bloggers specialise in writing about economics, innovation, and finances, whereas female bloggers prefer to discuss their lives and careers. AA studies rely on a limited dataset with a small number of writers and a large quantity of training data. In other circumstances, the problem can be directed using supervised ML approach testing on various stylometric features. Recent studies, however, revealed that traditional methods performed less efficiently concerning the accuracy and computation time generated new features, and interpreted the stylometric representation from the predictive model when applied to real-world applications, such as forensic investigation.

Koppel *et al.* [2009] approached the problems by using IR methods. First, they formed a training dataset through a user text of 500 words from 20,000 blogs. They evaluated the approach by 90% for training and testing with the remaining sample. Each snippet was then represented in three TF-IDF models based on 1,000 regular words and another based on style traits. An unlabelled writer is predicted by embracing a candidate whose known writing style has similar to the snippet. The three representations were evaluated; content classification reaches 56% higher, and the syntactic feature is far from benchmark values at 6%. Scalability is the issue for AA with several authors [Ruder *et al.* 2016a; Sari *et al.* 2017]. This study focused on identifying the writer of a document based on the stylometric representation from a character level using a deep neural network with thousands of potential candidates, even with limited training.

2.3.4 *Style change detection*

Imagine that a forensic investigator is provided with an online text snippet or document composed by an unknown number of writers [Singh *et al.* 2021], the problem of style change detection is to find the writing style of the document has changed. During the pre-trial investigation, something seemed strange; the writing style was inconsistent; however, the investigator cannot say what those inconsistencies are, and the court might want to determine the number of writers who contributed to this document under investigation, identifying who created a particular section of the document [Bevendorff *et al.* 2021; Muller 2019]. This task is related to the field of text forensic research for "*intrinsic plagiarism detection*" [Eissen and Stein 2006]. It could check the work of multiple writers in a certain area, such as how criminals use the Internet or social media to talk to one another, secretly, enabling online crimes. Other information sources include historical documents [Muller 2019].

Several works exist in the closely related field of intrinsic plagiarism detection, such as investigating whether an online text snippet or document is written by one or multiple writers [Eissen and Stein 2006; Hosseinia and Mukherjee; Stamatatos 2009a]. For example, Stamatatos [2009a] attempts to quantify the style variation within a document using character n-gram profiles and a style change function based on a dissimilarity measure originally proposed for writer identification.

Daks and Clark [2016]; Layton *et al.* [2013] researched clustering multiple documents, each compiled by a single writer; however, this is dissimilar to this study's undertaking because segments had to be established and grouped within a single document; it is unknown how the writer made the division. Akiva and Koppel [2013; 2012] employed unsupervised methods to identify various writer threads in multi-authored documents. The authors primarily focused on awarding credit for parts of texts to a known number of writers and left it to future research to determine the number of compilers.

2.4 LIMITATIONS OF THE EXISTING AUTHORSHIP METHOD

Despite the fact that various studies on AA or PAA have produced hopeful results, there are still substantial hurdles to this issue. Addressing these difficulties is critical for advancing the field of AA or PAA and ensuring its effectiveness. One of the most difficult tasks is devising solutions to overcome language barriers and cultural differences. Without overcoming these challenges, the full potential of AA or PAA may not be realised.

2.4.1 *Interpretation the results*

Interpreting the suggested method's results, whether from a regular machine or a deep learning model, is a significant challenge for a forensic investigator during a pre-trial investigation in court [Venčkauskas *et al.* 2015]. The investigator must demonstrate to the panel of judges the underlying factors that help identify the author of malicious or problematic material based on the author's writing style. Even though several studies have attempted to use various evaluation metrics and technologies to demonstrate that the model is accurate by demonstrating how the stylistic representation from the perspectives of lexical, syntactic, and semantic information [Sari *et al.* 2018a; 2017; Zhang *et al.* 2018c; Hu *et al.* 2020a] extracted from a collection of text is used to identify the author of a given text, the question of how the writing style helps quantify each among multiple authors has not been answered or explained [Venčkauskas *et al.* 2015; Hosseinia and Mukherjee]. As a result, additional research and breakthroughs in explainable AI are required to overcome these constraints and improve our comprehension of writing style analysis [Wu *et al.* 2021a; Zheng and Jin 2023].

2.4.2 *Small number of writers*

With the adoption of the Internet and the development of online texts via social media, emails, blogs, discussion forums, chat logs, and tweets have gained popularity. This has necessitated the development of more advanced natural language processing (NLP) and deep machine learning methods to accurately identify real authors from a larger pool of prospective writers. In doing so, the number of possible authors is typically very small, so it has to be increased to give more flexibility for the proposed methods to learn from known and unknown writing styles so as to determine the actual authorship of a suspected text. This can be achieved by incorporating a diverse range of texts from various sources, e. g.blogs, tweets, and LTS, and incorporating them as input into the training data. In another case, the actual author may be drawn from a group of writers because the numerous types of online texts may be generated by robots or ghostwriters based on the current world in which we find ourselves.

Text or words can be generated artificially by AI, so machine learning algorithms must be developed and used to distinguish whether the input text is an AI-made tool or was composed by humans before leveraging an NLP tool to extract and select linguistic elements in the text. As a result, advanced machine learning models must be developed in order to reliably identify authorship. This identification process can greatly enhance the accuracy and effectiveness of NLP applications. Allison and Guthrie [2008], for example, carried out a research investigation on identifying a writer using the Enron email corpus, which contains 4071 emails from 160 writers. The investigation used word n-grams as features and a multinomial probabilistic classifier, a hierarchical probabilistic classifier, and SVM as classifiers for writer prediction; the greatest accuracy was 87.1%.

Koppel *et al.* [2012] implemented a research experiment with 10,000 bloggers to determine the author of a text. Features used included content words, FWs, strings of non-alphanumeric characters, and strings of non-numerical characters. Identification was done using a similarity-based approach that went through several steps. As a result, the accuracy was 42%. These experiments demonstrate the effectiveness of machine learning algorithms in writer prediction.

Despite the relatively low accuracy of the features [Koppel *et al.* 2012], it still serves as a reminder of how difficult it is to accurately identify authors from text alone. Further research and improvement of feature selection methods may increase the accuracy of writer prediction models. Seroussi *et al.* [2014] highlighted that topic models frequently fail to capture the unique writing style and specific qualities of authors, making it difficult to accurately forecast writers based only on topic modelling. The author proposed that adding new features, such as syntactic patterns or semantic structures, could improve the accuracy of writer prediction models. However, the study found no common pattern for learning author writing style and no context for short or long documents.

2.4.3 *Training data size*

The length and volume of training texts, either obtained from the Internet or Twitter via the API, is a weakness for AA research that must be addressed. One complaint is a lack of continuity, which occurs when researchers do not make a long-term commitment to focus on a single subject [Rudman 1997]. Insufficient training text length and volume can lead to inadequate representation of the subject matter and hinder the accuracy of AA tasks. According to Ramnial *et al.* [2016a], it is crucial for AA and PAA researchers to include a sizable amount of written material with an appropriate number of words in each document. Additionally, [Sari 2018] demonstrates that inconsistencies in attribution datasets make it difficult to evaluate the performance of different systems. However, the use of language modelling methods has shown promising results in improving the accuracy and effectiveness of PAA or AA systems, as demonstrated by several studies [Fabien *et al.* 2020a; Jones *et al.* 2022].

Similarily, Argamon and Levitan [2005] measured the writer's name and country of citizenship of literary works gathered from a compilation of 20 works of literature written by eight individuals containing an approximate of a thousand phrases within each book by employing the sequential minimum optimisation (SMO) learning algorithm. Despite the fact that they obtained 99.9% classification accuracy for authors and 93.5% for nationalities, collections are characterised by a limited number of writings and relatively brief texts, which are predominant in a number of functional methods. However, these common methods often face challenges in accurately estimating the author and country of citizenship due to the limited amount of text available for analysis. Also, the study admits that the choice of default parameters in the SMO learning algorithm, which is a support vector machine (SVM), may not be the best for the dataset, which could affect the accuracy of the linearly weighted model. So, various ways to address the issue of shifting data scale and its impact on the comparison of commonly used terms and collocations were thus suggested. Again, [Dhar *et al.* 2022] puts forward a CNN-based author identification system for literary articles that makes use of visual features and a five-layer convolutional neural network. The system achieves a maximum accuracy of 93.58% on 1200 written works from 50 authors, surpassing the standard manually developed feature-based methods [Wu *et al.* 2021a] and established works on publicly available datasets. Although the proposed system is beneficial to society, it is unable to be adapted to handle a larger number of texts with multiple authors and different genres because there are no stylometric features, such as lexical, syntax, or semantics, but only feature visualisation, e. g.morphological features, which discriminate detrimentally against the input text by entirely gazing at the pattern of how they are woven rather than simply reading the written materials by observing their relevancy. This limitation poses a significant challenge as it fails to capture the nuances and intricacies of language usage across various texts. Without incorporating stylometric features, the system may struggle to accurately analyse and differentiate between texts from diverse authors and genres, limiting its effectiveness in handling a larger volume of such texts. Additionally, the absence of semantic analysis further hinders its ability to understand the meaning and context behind the written materials, further diminishing its adaptability in handling a wide range of texts.

The collaborative tasks encourage researchers and practitioners from various fields to collaborate and contribute their expertise in order to improve the quality and diversity of the datasets. By providing a platform for collective problem-solving, the PAN[15] evaluation lab aims to foster a sense of ownership and inclusivity among the community members, ultimately leading to more robust and reliable standard datasets. The PAN (Plagiarism, Authorship, and Social Software Misuse) evaluation, which is an annual conference covering many different attribution tasks, beginning with standard AA and progressing to more advanced tasks like author clustering and obfuscation, launched the campaign. CLEF [16] is also known as the Conference and Labs of the Evaluation Forum. In a paper submitted in 2016 by [Potthast *et al.* 2016], the authors report on the first large scale evaluation of author obfuscation methods implemented to interfere with authorship verification and report the system's performance. All of the systems submitted for authorship verification and obfuscation reached an average of 47%, and more of the published work wrongly attributed a set of texts to a different or multiple authors when the system would have come to a different conclusion if one of the texts had not been marked as obfuscated. Similarly, Tschuggnall *et al.* [2017] offered an assessment of style breach detection and clustering in order to quantify multi-authored works with their individual writing styles. The submitted models were evaluated using the PAN 2016 dataset, which includes English, two non-English languages, and two different genres. Three methods for detecting style violations were published, all of which used stylometric features and word dictionary databases in conjunction with different distance metrics or adopted a neural network similar to the word embedding method. Even though the proposed method outperformed the baseline method, only one of them slightly outperformed it, which is based on random guesses. Surprisingly, the method only considers the end of paragraphs as possible features, not whether the paragraphs exist to extract writing style along with a fixed length of the text that goes into the model. This limitation suggests that the proposed method may not fully capture the nuances and intricacies of writing style. Therefore, further research is needed to explore additional features and improve the accuracy of the model.

Despite CLEF's attempts to diversify the task through PAN, the majority of the released datasets are small. For example, only 100 training cases were provided for each language in the PAN 2015 authorship verification task. Reduced data limits participants' utilisation of more advanced methods, which require enormous quantities of training data, so most methods proposed by the participants only show the effectiveness of the feature, not how the feature contributes to detecting or identifying the author of the given text. In the majority of cases, the authors used convolutional MLA, such as SVM and stylometric feature metrics Section 2.2. In general, some studies concentrate on feature discovery and data processing in order to figure out the most effective writing style from a given text or input [Vosoughi *et al.* 2016; Neal *et al.* 2017b]. However, these have not helped to understand the linguistic writing pattern in a given literary document, not to mention the short text posted on social media that is short in character, and the words in the text are synonymous with abbreviations and acronyms. Therefore, Bagnall [2015] implemented a character-based language model that predicts the next character in an ordered sequence based on the specific characters that came before it in the sequence. The proposed model is based on recurrent neural networks (RNN) and validated on the AV dataset. The author also stated that the high computational cost was due to the complexity of the RNN architecture and the large number of parameters involved. Despite this limitation, the model's results demonstrated its effectiveness in capturing unique linguistic writing patterns and achieving accurate results for AV datasets. Aborisade and Anwar [2018] incorporated logistic regression (LR) and naive Bayes (NB) to categorise tweet authors, and the study achieved a higher accuracy (91.1%). In the same way, Tanaka and Jin [2014] used characters with n-grams to identify email authors from a cell phone conversation message.

15 https://pan.webis.de/
16 https://www.clef-initiative.eu/

Because of privacy concerns, the researchers didn't say how they got their hands on the dataset, but the model's F1-score was 0.92% using RF.

In general, the training dataset is always too small to make a meaningful prediction, and using a literary document for the AA problem as a downstream task on PAA will not work because the writing structures of both are different. In this study, we crawled social media datasets, particularly Twitter, with specific # tags of authors or influential people in society for a period of one year and mined the publicly available dataset on an AA task to evaluate our proposed regularised neural network model. Our approach aimed to address the limitations of previous studies by utilising a larger and more diverse dataset from social media platforms. By focusing on specific hashtags related to authors or influential individuals, we ensured that our training dataset captured a wide range of writing styles and language variations. This allowed us to effectively evaluate the performance of our regularised neural network model on the AA task.

2.5 CHAPTER SUMMARY

In Chapter 2, we began by reviewing online snippets of text on social media, their underlying characteristics, and the attribution techniques used previously based on stylistics and content-based features in conjunction with statistical data and machine learning methods. We then explored the concept of author profiling, which involves analysing various features of an author such as age, gender, and personality traits to determine their identity. Additionally, we discussed the advantages of stylistic-based methods, which focus on linguistic patterns and writing style to identify authorship. However, it is important to note that these methods may not always be accurate due to the potential for deception or multiple authors using similar writing styles. Second, we investigate the basic feature sets, the challenges of extracting these traits, and the methodology for determining the writing style of anonymous or disputed documents during a pre-trail investigation. Finally, we investigated the potential applications of authorship attribution in a variety of contexts, including closed set authorship attribution, authorship verification, and author profiling, as well as the benefits and drawbacks of each method. However, one of the issues with post-authorship attribution (PAA) or authorship attribution (AA) that has not been thoroughly investigated with natural language processing, deep learning models, and from a computational linguistic perspective is understanding how the model learns the feature and which part of the stylistic features contribute more or are more informative in determining the author of anonymous text or documents dynamically and can be visualised. This gap hinders the ability to fully understand the inner workings of these models and their limitations. Additionally, without this knowledge, it becomes challenging to interpret and trust the results produced by these models in real-world scenarios. Therefore, further investigation is needed to shed light on the learning process of these models and identify the most influential stylistic features for accurate authorship attribution.

Part III

AUTOMATING EXPLORATORY STATISTICAL ANALYSIS WITH MACHINE LEARNING

Exploratory statistical analysis (EDA), like exploratory data analysis (EDA), is the process of evaluating our dataset collection using statistical and visualisation approaches to highlight relevant features of the data for future study. This entails analysing the dataset from a variety of perspectives, interpreting it, and summarising it without making any assumptions about its contents. In this thesis, EDA is an important step to conduct before moving on to statistical modelling or machine learning to confirm that the data is what it claims to be and that there are no visible flaws. It should be incorporated into all data science efforts. EDA aids in the identification of patterns, anomalies, and correlations within the data, which can provide valuable insights for decision-making. EDA is an essential step in the data analysis process.

3

AUTOMATING EXPLORATORY STATISTICAL ANALYSIS WITH MACHINE LEARNING

In Chapter 3, we engage in exploratory statistical analysis (ESA) to look at the descriptive characteristics of the data that can be visualised and summed up statistically in order to gain insights into its structure, patterns, and potential issues. Exploratory statistical analysis (ESA), which can also be called exploratory data analysis (EDA) from a data science perspective, could aid in recognising abnormalities and contradictions in the dataset, allowing forensic investigators to better understand patterns within the data, such as detecting outliers and unexpected events, and identify intriguing connections between features. Visualising the distribution of text in the dataset can also help identify the issues of sentiment and phylogenetic within the data, which could shed light on cultural and historical circumstances. Furthermore, the use of ESA can aid in analysing literary styles, which can help to understand the influence of different authors on one another and the trends in writing styles that overlapped over time. Additionally, the exploration of ESA can provide valuable insights into the evolution of language and its impact on society. Despite having four datasets to evaluate the proposed model in Chapter 4, Chapter 5, and Chapter 6, we only experimented with ESA on the CCAT50 and Twitter datasets. The two datasets use the same nomenclature as subjects in relation to OTS and LTS, which provides useful insights into the interdependence of communication and societal change via social media. By focusing on these two datasets, we can gain a comprehensive understanding of the relationship between language evolution and societal change in the context of online communication. Additionally, analyzing these datasets allows us to draw meaningful conclusions about the impact of social media on language trends and writing styles.

3.1 INTRODUCTION

People generally agree that social media, like email, blog posts, forum posts, or chat logs via an online text snippet (OTS), has a big impact on our lives. It gives us a way to stay in touch, no matter how far apart we are, and a lot of people spend their days on the platform [Modupe *et al.* 2022]. Most of the OTS are short text messages that are around 280 characters or less. In this particular scenario, OTS can generally be collected via an API or downloaded from a repository into an array (e.g., a spreadsheet or database). and typically stored with one row per experimental subject and one column per subject identifier. Each of these collected texts could have numerical values for a specific quantitative variable or categorical variable level. On the other hand, authorship analysts or forensic investigators are trained or rely on the fact that each individual has a distinct writing pattern (e. g. stylistically expressive elements), which can be extracted using exploratory statistical analysis (ESA) and visualized for predicting users' interest preferences in the data and used as a supplement to language-based evidence in a court of law during the pre-trial investigation [Venčkauskas *et al.* 2015]. ESA can assist in finding various linguistic and stylistic characteristics in the text, such as discriminating words and phrases, as well as writing style.

Typically, ESA is done by interactively applying various statistical operations (such as filtering, aggregation, and visualization) to find interesting common patterns in a text and determine whether the features or terms were written by the author or someone else [Milo and Somech 2020]. In addition, ESA can be used to gain deeper insights into the overall content of a text by studying the frequency and patterns of words used in context. For example, discovering an intriguing pattern in a text could assist law enforcement in being proactive in understanding the social concern about the assault on US Congress members, particularly

those who objected to the election's results in 2021 [Grimmer 2010; Monroe *et al.* 2008; M. Otala *et al.* 2021]. Furthermore, the language in a text or OTS can provide information about the author's writing style (e. g.personality, gender and age) [Schwartz *et al.* 2013a] and online expressions of consumer sentiment about restaurants [Jurafsky *et al.* 2014]. This kind of insight can be particularly helpful in understanding public sentiment and how it changes over time. Therefore, a wide range of visualisation methods, such as KrackPlot[1] [Krackhardt *et al.* 1994], Pajek[2] [Dohleman 2006], UCINET[3] [Apostolato 2013], and Visone [Baur *et al.* 2002], have been developed to assist analysts in understanding the pattern in the dataset. In addition, Dimitriadou *et al.* [2016]; Kraska [2018]; Milo and Somech [2018] facilitate the process of using multiple dimensions for automating ESA or better known as exploratory data analysis (EDA) but it is time-consuming to process and evaluate, especially for non-expert users. First, simplified EDA interfaces (e. g. [Kraska 2018], Tableau[4]) allow non-programmers to effectively explore datasets without knowing scripting languages or SQL. Secondly, Crotty *et al.* [2016]; Elbaz *et al.* [2020]; Kamat *et al.* [2014] proposed a devised solution to improve the interactivity of EDA, by reducing the running times of exploratory operations, and displaying preliminary sketches of their results. Although all these studies substantially speed up and simplify the exploration process, ESA is still a challenging process because the user is still in the "driver's seat" and must choose which exploratory operation to use at each stage of the exploratory process [Milo and Somech 2020]. This decision-making process requires an in-depth knowledge of both the data and the problem being explored, as well as an understanding of which operations are most likely to yield meaningful results. These methods fail to understand the relationships among the explanatory variables, and a large number of overlapping links makes judging connectivity difficult [Perer and Shneiderman 2008]. Identifying the direction and approximate size of relationships between explanatory and outcome variables, on the other hand, remains difficult. Again, these methods are limited to highlighting discriminating words that can be found in ranked lists of words, word clouds, word bubbles, and word-based scatter plots [Dicle and Dicle 2018] to find an interesting phenomenon or discriminating word or phrase in the data. Even with the word cloud that legibly displays term labels in scatter plots [Cliche *et al.* 2017], it is difficult to compare the relative frequencies of two variables (or terms) in the data.

In Chapter 3, we present a collection of highly diverse post-authorship attribution datasets that better generalize evaluation and ESA to discover, explore, and detect phenomena in each of the datasets considered in this thesis. The aim is to use ESA with a query-based visualization to show how each writing pattern in each text with similar embedding differs between texts by the same authors. Thus, the unique contributions in Chapter 3 are as follows:

1. We employed statistical and data visualisation methods to help identify deeper relationships, patterns, common subjects, and emotions in each text, revealing vital information and influencing text structure.

2. To conceptualise the phenomena in the data, we present exploratory statistical analysis and visualizations.

3. To gain deeper insights into the entire corpus to reduce the number of outliers under different assumptions.

The remainder of Chapter 3 is structured as follows. In Section 3.1, we provide the introductory message and Section 3.2 we provide a comprehensive overview of the dataset. In Section 3.3, we look at the univariate graphical ESA and visualization methods for analyzing

1 https://www.isi.edu/~blythe/KP/
2 http://mrvar.fdv.uni-lj.si/pajek/
3 http://www.analytictech.com/archive/ucinet.htm
4 https://www.tableau.com

interesting pattern in the dataset. Finally, in Section 3.4, we present the conclusion and future work on fully automating ESA.

3.2 AUTHORSHIP ATTRIBUTION DATASET

A variety of data sources have been examined for the purposes of post-authorship analysis (PAA) and authorship attribution (AA). In this Chapter 3, we experimented with three previously used corpora, including CCAT50, IMDb62, and Blog50, as well as the most recently collected Twitter datasets using the API outlined in Section 2.1.1. The datasets provide evidence of feature representations, which are implemented to develop writing styles, as well as deep learning methods for predicting the writer. To accomplish this, the quality of the dataset must be guaranteed, specifically [Rudman 2012; Neal *et al.* 2017a; Sari 2018]. Other possibilities include deleting non-alphabetic characters and spaces, converting capital letters to lowercase, stemming or lemmatizing, encoding, and human writing.

Additionally, analysing the frequency of certain words or phrases used by the writer can also provide insights into their writing style. Furthermore, comparing the dataset with known writing styles or authors can help identify similarities or patterns that can contribute to understanding the writer's unique style. This analysis can be useful for various purposes, such as determining authorship or detecting plagiarism. These are critical techniques for extracting a writer's writing style from a collection of datasets, resulting in important insights and applications.

Table 3 provides an overview of fundamental statistics and a broad range of influencing factors to help comprehend the major qualities and visualise trends in the datasets. The dataset has a authors, d documents, and w words. Furthermore, $\bar{x}$ assessed the length of text in the documents for each dataset, and μ measured the average number of documents for each dataset. The whole dataset is in English, and all non-English content has been eliminated to ensure the accuracy of the analysis.

Table 3: Statistics of the datasets

Dataset	Text type	a	d	w	$\lvert\bar{x}\rvert$	μ
CCAT50	Industrial/Financial news	50	5,000	1.254M	502	50
IMDb62	Movie reviews	62	61,986	16,904M	341	800
Blog50	Blog articles	19,000	681,288	167K	562	35
Twitter	Social media comments	1,391	1,109,964	1,259M	532	35

Although Table 3 shows that the number of authors for Twitter, Blog, and IMDb is greater than that for the CCAT50 dataset, in this thesis, we chose the most prominent 50 authors. This strategy proves to be more powerful or superior to the benchmark proposed frameworks in Chapter 4, Chapter 5 and Chapter 6 compared to SOTA methods, boosting the reliability of the results and datasets. Furthermore, by selecting the most prominent 50 authors, we ensure a diverse range of perspectives and writing styles, which enhances the generalisability of our findings. Additionally, this approach allows us to focus on the most influential personalities in the field, providing more impactful insights for future research and applications.

3.2.1 *CCAT50 dataset*

The CCAT50 dataset[5] is a subset of Reuters Corpus Volume 1 (RCV1) [Russell-Rose *et al.* 2002; Lewis *et al.* 2004]. The CCAT50 dataset contains over 800,000 manually categorised newswire stories written in English that cover a wide range of topics, including society, politics, business, financial, industry and corporate news articles. The corpus is divided into CCAT10 and CCAT50, respectively. First, the CCAT10 was divided into 50 training and 50 test texts per author. In our experiments, we partition the datasets into training and testing as provided by prior work [Stamatatos 2013; Sari 2018] to measure the performance accuracy of the proposed model. The CCAT50 corpus, on the other hand, is a larger version of the CCAT10. In total, there are 5,000 documents from 50 authors. As for CCAT10, for each of the authors, there are 50 training documents and 50 test documents. A sample text from the CCAT10 and CCAT50 datasets is shown in Figure 6 after preprocessing by removing unnecessary spaces, tags, extra spaces, stop words, and non-ASCII characters.

"the watchdogs of china ruling communist party have ordered members to obey call from party chief jiang zemin for ideological orthodoxy or face disciplinary punishment state media said on saturday the work style of the ruling party is matter of life and death the people daily quoted decision by the central commission for discipline inspection of the communist party of china"

Figure 6: A snippet from CCAT50 datasets

3.2.2 *IMDb62 dataset*

The IMDb62 dataset is a popular Internet Movie Database (IMDb)[6] collected in May 2009 and introduced by [Seroussi *et al.* 2010]. The total number of reviews is 62,000, with 17,550 messages from 62 prolific users. Each user wrote 1,000 reviews of movies they liked based on their favourite TV shows, documentaries, poems, and other entertainment news. Among the datasets we used in our experiments, IMDb62 has the most authors and documents. Compared to the CCAT50 dataset, the style elements in IMDb62 are less formal and shorter (see Figure 7). IMDb62 represents an interesting dataset for researchers and practitioners working in the areas of natural language processing, text classification, and sentiment analysis.

"few films capture true stories as well as the elephant man. although there were of course certain artistic liberties drawn as with any film based on actual events david lynch respects joseph merrick enough not to present him as a monstrous caricature of a man you may be wondering why i do not refer to joseph Merrick as john Merrick. his real name was actually joseph but was changed for the film as its screenplay was based on correspondence letters between dr frederick treves and joseph for some reason he altered Josephs name in the letters to john at some point in time no ones sure why john or joseph merrick had a rare disease called proteus disease a congenital condition made him appear hideously deformed and unrecognizable as a man...."

Figure 7: A snippet from IMDb62 datasets

3.2.3 *Blog dataset*

The Blog dataset is a truncation of the expression "weblog" and the most extensive dataset, containing 678,161 blog posts by 19,320 authors collected from August 2004 and before [7].

5 https://trec.nist.gov/data/reuters/reuters.html
6 http://{www.imdb.com}
7 blogger.com

The dataset was introduced as part of a study on the effects of age and gender on blogging by Schler *et al.* [2006a]. The blog dataset provides an invaluable resource for understanding the dynamics of online behaviour, allowing researchers to analyse how people use blogs to create relationships and express their ideas. But many bloggers must ensure that every topic discussed in the blog is something that all authors are interested in, and Koppel *et al.* [2011] used a different blog consisting of 10,240 authors in their work on authorship attribution with many candidate authors. An average of 35 posts were collected per author, and no preprocessing or filtering was applied to the corpus. It is worth mentioning that this dataset is the one containing the shortest texts on average (79 tokens for the top 5 authors). Many of the blog posts collected were replies to existing blog posts or short articles. For the Blog dataset, the texts are short, and the number of texts per author is smaller than for CCAT50, with less variability than for IMDb62. Many of the blog posts collected were replies to existing blog posts or short articles. As a result, in our experience, we considered the top 10, 20, and 100 authors with the largest number of texts. Thus, Figure 8 shows a sample of the text from the blog.

"wearing tool band tee dicky sock listening too jenny wrap a present thinking today is a good day im at jenny house right now im staying the night it going to be great um in english i have this huge assignment thats going to take me forever and theyre only going to give me a week to do it i have to write three 500 word essay about the pain of procrastination read an entire book tuesday with morrie and come up with a presentation about a clockwork orange it gonna suck well i dont know what else to bitch about bye kid"

Figure 8: A snippet from Blog datasets

3.2.4 *Twiiter dataset*

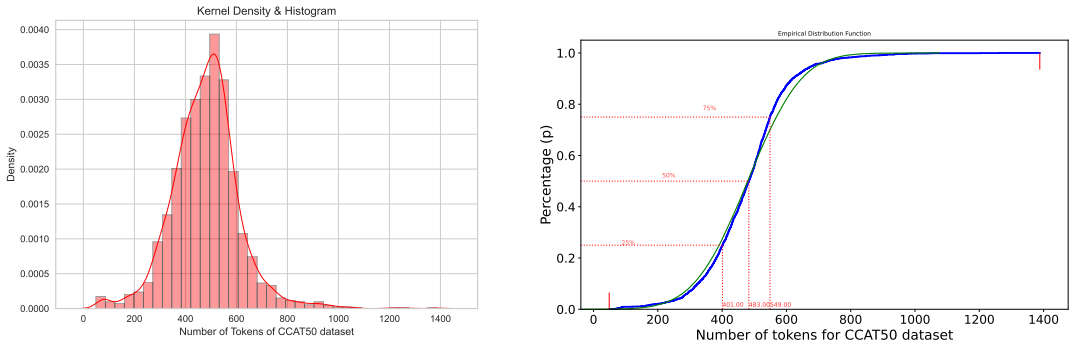
For the Twitter dataset, we implement the same instance as in [Rocha *et al.* 2016] by simply using tweet IDs to collect a new and larger dataset rather than entire tweets, which is one of the constraints with the Twitter API or searching API [Cutler *et al.* 2020]. The API's rate restriction is 450 requests per 15 minutes, with a maximum of 100 tweets per request. We aggregate the user to a list of 4,391 celebrities such as columnists, politicians, technologists, artists, and influencers on social media in 68 categories covering everything from politics, social unrest, and tech to arts and culture. We gathered over 1 million tweets from over 1,000 influential users, and the new dataset has allowed us to gain deeper insights into the dynamics of social media conversations, enabling us to study topics such as opinion formation and consensus as well as public responses to events in real time. For the Twitter dataset, we combine all the tweets of a single user into a single document and treat each tweet as an individual sentence. To proceed with the dataset in our experiment, we split the datasets by sampling 60% of each author's documents into a training set, 20% for validation, and rendering the remainder for testing over 10-fold cross-validation as used in most AA tasks.

3.3 EXPLORATORY STATISTICAL ANALYSIS AND VISUALISATIONS

This analysis aims to find interesting patterns in the dataset that could be used to visualize and quantify linguistic variation in the data. First, we employed univariate graphical analysis for ESA, a qualitative method that involves a degree of subjective analysis to help gain insight into interesting patterns in the data that may be linked to socio-economic factors. Secondly, we look at using scatter text methods [Kessler 2017] to analyze how the text (e. g.lexical features) in the dataset are distributed based on the sentiment scores.

3.3.1 *Histograms*

In exploring the use of exploratory statistical analysis (ESA), we use a histogram to demonstrate the frequency variation of the data. It helps identify data distribution, outliers, skewness, and other shape aspects. Therefore, we create the histograms by dividing the data points into a variety of intervals called bins, where the height of each bar corresponds to the frequency[8] or relative frequency of data falling within that bin. This allows for a quick and intuitive understanding of the underlying data patterns. As an example, Figure 9 a shows the statistical distribution of a 1000-item sample from the CCAT50 dataset. It also shows how document lengths in the text change based on the size of lexical tokens in the CCAT50 dataset, with a density value of 0.0035. While the Figure 9 b represents the overall cumulative distribution function (CDF), it shows that 75% of the documents have an average length of approximately 500 tokens. This visual representation aids in comprehending the spread and central tendency of the textual content in the CCAT50 dataset.



(a) Histogram (b) CDF

Figure 9: Histogram and CDF on the CCAT50 datasets.

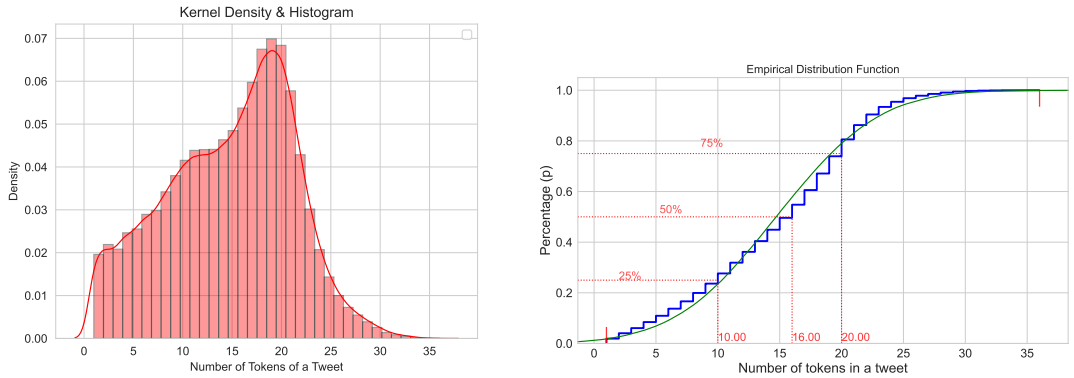
In the Twitter50 dataset, Figure 10 a shows the distribution histogram and CDF for tweet length (textual content), represented by a vocabulary symbol. Thus, 75% of tweets include approximately 20 tokens, indicating a density value of 0.06 (e.g., height). According to Figure 10 b, the CDF reflects 75% of the length of a Twitter tweet, which is about 20 tokens or 20 words. This emphasises how difficult it is to communicate important information in the constrained space that social media platforms offer. Additionally, the brevity of these messages may lead to misunderstandings or misinterpretations among users.

3.3.2 *Quantile-normal plots*

The quantile-normal, or QN plot, or, more generally, the quantile-quantile, or QQ plot, is a graphical method for determining if two data sets come from populations with a common distribution. It is the most complicated univariate graphical ESA method we employed on the data to find interesting patterns. In many statistical tests, there is an assumption that the outcome of a fixed set of values from data will approximately follow a normal distribution, but it has been proven that the assumption is grossly violated, and the QN plot is useful to measure the confidence interval using a p-value.

The example shown in Figure 11 shows many points in a row that are on the same side of the line (rather than just bouncing around to either side), which suggests that there is a real

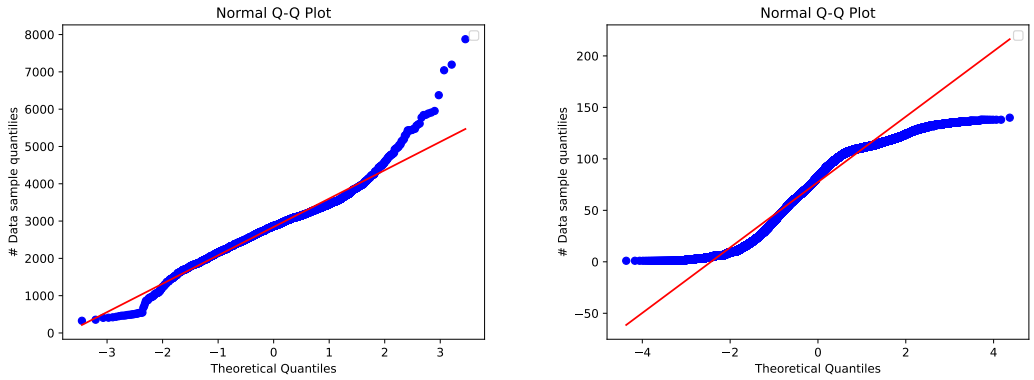
8 e. g. count or proportional of total count in each data point for a range of values



(a) Histogram (b) CDF

Figure 10: Histogram and CDF on the Twitter datasets.

(non- random) deviation from normality. In each area, we see how the observed values deviate from normality. In each area, we see how the observed values deviate from what is expected in the histogram, e. g.in which the "x" (observed value) appears to have moved relative to the "*perfect normal*" line. Here we observe data points or values (e. g.writing style) that are too low from a certain sample of the data at the beginning but pulled asymmetrically towards higher values, which indicates positive skew (e. g.normal distribution) at a certain area in the texts.



(a) CCAT50 (b) Twitter

Figure 11: A quantile-normal plot.

In Figure 11 (a) there is multiple point that is outside (or off) the reference line, i. e. shifted to the right of where it should be and this pattern shows nearly Gaussian data with one "high outlier". Similarly, Figure 11 (b) looks a bit similar to the "skew left" pattern, but the most extreme points tend to return to the reference line (e. g.bi-modal data).

3.3.3 *Correlation and covariance*

In analyzing the concept of discovery, exploration, and empirical detection of basic phenomena in data using ESA, covariance and/or sample correlation are used, which correspond to a measure of how much two variables "co-vary", e. g.how much (and in what direction) should we expect one variable to change when the other changes. By understanding the correlation between variables, it is possible to determine which independent variable is affecting the dependent variable. Thus, positive covariance values suggest that when one measurement is above the mean, the other will probably also be above the mean, and vice versa. negative

covariances suggest that when one variable is above its mean, the other is below it. Thus, Figure 12 shows the correlation matrix on how the phenomena, e. g.style features, in the Twitter dataset are structured. The positive and negative correlation values show a strong link between how an author's writing is structured and how it is written in terms of literary elements such as sentiment, punctuation, formality, excitement, etc. For example, there is a strong correlation of around 88% between the length of the character sequence and the frequency of the number of words in the Twitter dataset, which demonstrates the numerical relationship. Similarly, there is a 0.51% difference in calculating the length relating to POS tagging in the text compared to when the content is tagged as adjective or the average number of words, unique ratio, and punctuation marks (syntactic), the average number of words (lexical), and the content length (semantic). This correlation further reinforces the notion that writing style and its underlying meaning are intricately connected.

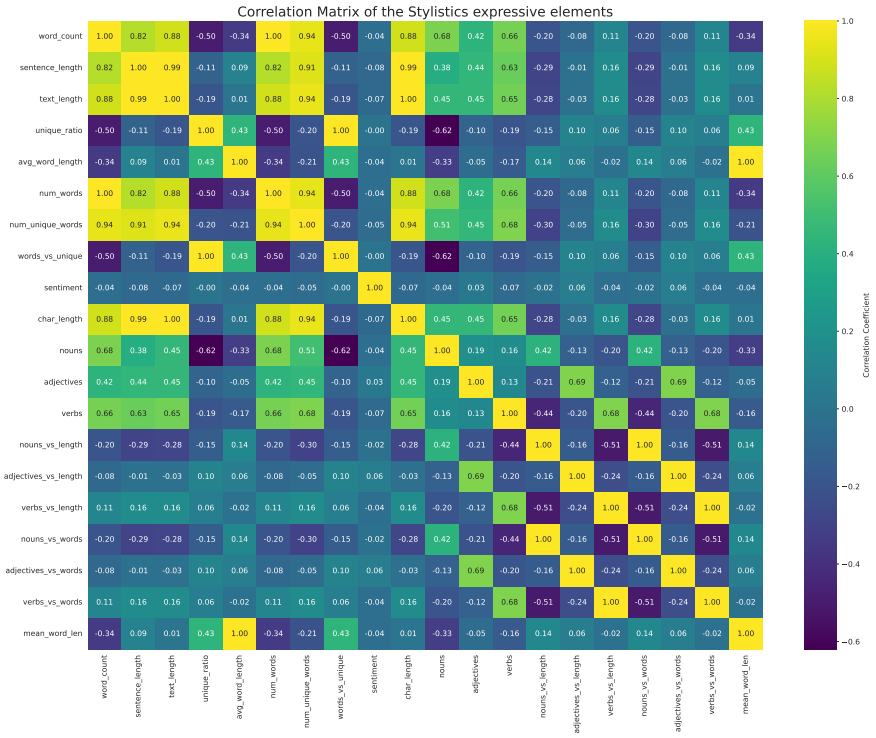
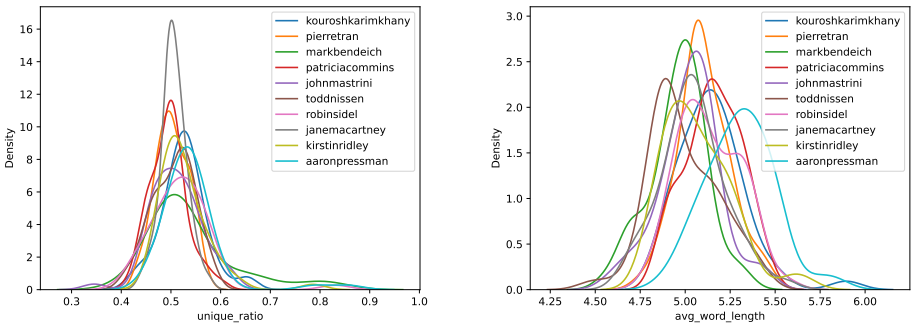


Figure 12: Sample covariance heatmap for Twitter dataset.

Furthermore, we found a strong correlation between word length and the number of unique words used by each author in the CCAT50 dataset by building a list of words and removing duplicates to determine the distribution of unique stylistic elements (e. g.uncommon words) (see Figure 13). It seems like the usage of unique words is about the same (about 50% of words are unique), though the distribution varies a bit in size compared to the actual words used (e. g.longer words for each author in Figure 13).

The same result is shown in Figure 14 for the CCAT50 dataset. The yellow and blue color coding of the cells makes it easy to identify the linear correlation, e.g., the relationships between writing style and stylistic phenomena in the text. The cells that appear in yellow depict a positive linear correlation, and those that are coloured blue show a negative linear correlation. Additionally, the numerical value between -1 and 1 indicates the strength of the correlation. As shown in Figure 14, there is a strong correlation of 95% between the unique number of words and the length in the text compared to when the POS tagged of a word is noun and the length in the text with 1% relationship. We removed duplicates from a list of



(a) Unique word ratio (b) Average word used/authors

Figure 13: Unique and average distribution in the CCAT50 datasets.

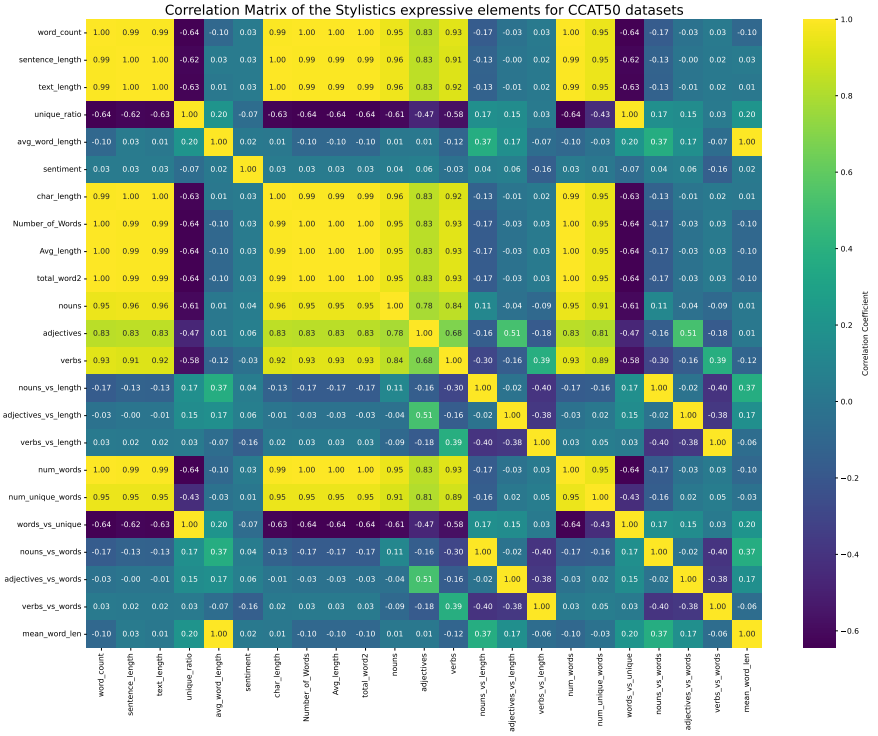
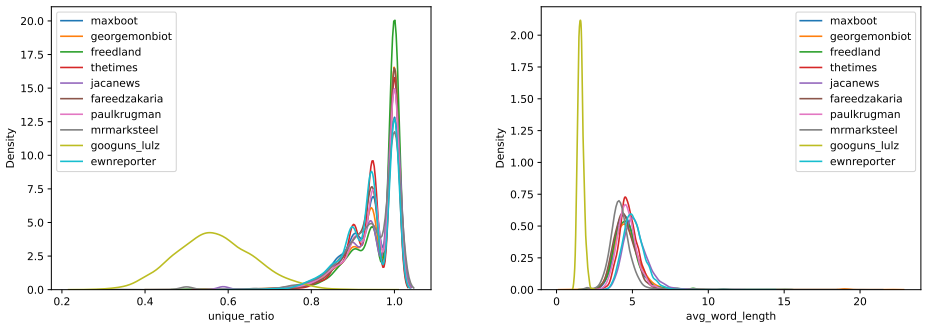


Figure 14: Sample covariance heatmap for CCAT50 dataset.

top K-authors in Twitter datasets to examine the distribution of distinct words. As shown in Figure 15, most authors have an average word length of 50%, and 75% of them contribute uniquely to the sample text's length. The results show that most authors in this dataset use words of similar length and uniqueness.

3.3.4 *Sentiment factors*

The VADER model [Hutto and Gilbert 2014] is used to express the relationship between two opposing concepts within text, such as finding emotionally charged issues in the distribution of the dataset. We implement this by extracting grammatical and lexical features to determine if the text emphasises or expresses sentiment intensity (e. g. sentiment scores). This analysis



(a) Unique word ratio (b) Average word used/authors

Figure 15: Unique and average distribution in the Twitter datasets.

provided valuable information about the contextual sparseness of the authors' sentiments and helped us understand their overall sentiment expression. Listing 1 is a snippet of implementation code called "polarity scores" that takes a piece of text as input and generates a dictionary containing the sentiment scores for the text. The dictionary contains four keys: negative, neural, positive, and compound, where a negative sentiment score is between 0 and 1. For the CCAT50 dataset, we achieve a negative score of 9.7%, a neural score of 74%, a positive score of 1.63%, and a compound score of 42.15%, as described in Listing 1. This is the sum of the positive, negative, and neutral (e. g.valence score of each word in the text) scores. Thus, the compound scores are normalized between -1 (the most negative) and 1 (the most positive) to give a good overall classification and update the dataset with the sentiment scores. On top of that, Figure 16 shows the sentiment scores for the top-10 authors to infer how each individual author differs in their writing.

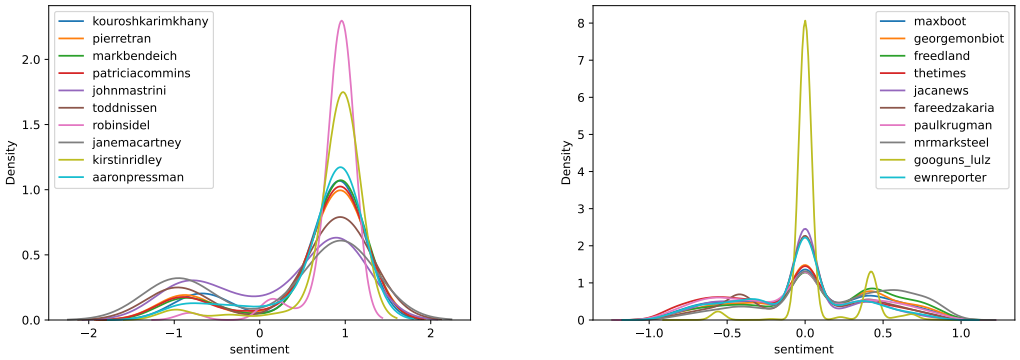
Listing 1: Sentiment factor analysis

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sid = SentimentIntensityAnalyzer()
# we test with the sample texts
print(sid.polarity_scores(CCAT50["clean_text"])
\\
print(------------------------------)
{"neg": 0.097, "neu": 0.74, "pos": 0.163, "compound": 0.4215}
print(------------------------------)
```

For the Twitter dataset, with the top-K author selected from the set of candidate authors, "*googuns_lulz*" has more neural expression (80% density) in their writing compared to other sentiment scores such as positive and negative (40% and 10%, respectively). In contrast to the other authors, the probability density is significantly lower and and author "*robinsidel*" has a higher positive density compared to the others for the CCAT50 datasets (see Figure 16(a)).

3.3.5 *Interactive tools for ESA visualization*

In this section, we present a simple mathematical notation for visualising and explaining the ESA as a collection of terms from a dataset corpus that can be assigned to one of two categories on a two-dimensional plane using a scatter-plot [Urribarri and Castro 2017]. The visualisation is a scatter-text [Kessler 2017] plot of uni-grams and bi-grams. The scatter-text plot allows for the identification of key terms that are most representative of each category, and we used it in this thesis to explore the relationship between these terms and the categories they represent. It is applicable to a wide range of datasets and provides valuable insights into the underlying structure of the data for improving the extraction of stylistic representation for PAA or AA



(a) CCAT50 (b) Twitter

Figure 16: Sentiment factors.

tasks.

To find an intriguing phenomenon in the data and visualise it in an interactive manner, we consider a corpus of input text D with disjoint subsets P and Q such that:

$$P \cup Q \equiv D \tag{1}$$

where $\cup^*$ represents the disjoint subset of P and Q.

In addition, we let $\psi^T(t, D)$ be the number of times term occurs in D, $\psi^T(t, P)$ be the number of times t occurs in P. We let $\psi^D(t, P)$ refer to number of documents/texts in P containing t. Mathematically we let $t_{(ij)}$ be the j-th word in t_i where $j \in \{1, 2\}$. So, the parameter of ψ may be ψ^T or ψ^D as a features representation from the notion of term frequency-inverse document frequency (tf-idf) as follows.

$$P[t_i] = \frac{\psi(t_i, D)}{\sum_{t \in D \wedge |t| \equiv |t_i|} \psi(t, D)} \tag{2}$$

The objective of employing tf-idf to construct the set of terms is to include the visualisation V in the process and in order to model the inherent subjectivity in a natural language to understand and analyse the phenomenon in the writing style of each author in the given data, either as a word or phrase. Thus, we follow Schwartz *et al.* [2013a] approach and use a point-wise mutual information (PMI) score to filter two-word sequence of words (or bi-grams) that do not occur far more frequently than expected in the documents.

So, we define the PMI as:

$$\text{PMI}_{(t_i)} = \log \frac{P[t_i]}{\prod_{t_{(ij)} \in t_i} P[T_{ij}]} \tag{3}$$

where the minimum value of PMI accepted is $\boldsymbol{p}$ and we define V as

$$\{t | \psi(t, D) \geqslant \boldsymbol{m} \wedge (|t| \equiv 1 \vee \text{PMI}(t) > \boldsymbol{p})\} \tag{4}$$

Following the approach in [Kessler 2017], we define the term t's coordinates on the scatterplot [Urribarri and Castro 2017] to be (x_t^P, x_t^Q), where P and Q are the two documents categories (or labels). x_t^K is proportionate to $\psi(t_i, K)$, and many of the terms will have identical $\psi(t_i, K)$ values (as word or phrases) and to break the ties, meaning the word that appears last alphabetically will have a larger x_t^K.

In addition, we define g_t^K such that $t \in V$ and $G \in \{P, Q\}$ as the ranking of $\psi(t, K)$, arrange (or sorted) in ascending order, where the ties are broken by terms' alphabetical position [Kessler 2017]. Hence, this allow to define the coordination position in the scatter plot as

$$x_t^K = \frac{g_t^K}{\text{argmax}\, g_K} \tag{5}$$

where score values of x is limited to $[0, 1]$, ensuring that both axes on the scatterplot are scalable in the same way (or identical). This actually helps to keep the graph from becoming lopsided toward the corpus that had a larger number of terms.[9] This is because ordinal ranks give more weight to important words and allow for better visualization of meaningful data patterns. Additionally, log frequency can sometimes obscure important information by compressing the axis and making it difficult to distinguish between different data points. The Figure 17, Figure 18 and Figure 19 were designed with $\boldsymbol{m} = 4$, $\boldsymbol{p} = 7$ and $\psi = \psi^{\mathrm{T}}$.

3.3.5.1 *Phrase associations*

In this section, we attempt to identify domain-indicative lexical items e. g. a variety of words or phrases in the document) and be able to visualise them. At first, we employed a graph centrality algorithm, the TextRank package in Python [Mihalcea and Tarau 2004] to extract the scores of the most prominent phrases in the document on the basis of the sentiment score using VADER [Hutto and Gilbert 2014]. Furthermore, we normalise the corpus using spaCy libraries in Python [10] [Al Omran and Treude 2017], where we parse each document as opposed to normal tokenization separated by whitespace (see Listing 2) [Babur *et al.* 2022].

Listing 2: TextRank algorithm for prominent phrases

```
import pytextrank, spacy
import scattertext as st
nlp2 = spacy.load("en_core_web_sm")
nlp2.add_pipe("textrank")
corpus2 = st.CorpusFromParsedDocuments(tweet,
                        category_col='vader_prediction',
                        parsed_col='parse3',
                        feats_from_spacy_doc=st.PyTextRankPhrases()
                    ).build().compact
                    (AssociationCompactor(2000,
                        use_non_text_features=True))
```

The phrases generated in Listing 1 are reduced to 2000 and treated as non-textual features because the document scores are not word counts. The terms, on the other hand, are name entities rather than frequency counts with eigencentrality, which measures the node's influence scores based on the TextRank algorithms [Mihalcea and Tarau 2004].

Table 4 plots the interactive graph using a generalized scatter plot [Keim *et al.* 2010] based on the dense ranks of the scores.

Because the aggregate TextRank scores in Table 4 cannot be interpreted, so we use the *meta_data* description function in Listing 3 to visualize the most used terms that are associated with each author in the text based on the per–class ranking of each score.

Listing 3: Meta Description function

9 *Ordinal ranks are preferred over log frequency due to uninteresting stop-words.*
10 https://spacy.io/universe/project/spacy-pytextrank

Table 4: TextRank algorithm scores for Twitter datasets.

Terms	Positive	Natural	Negative
the en	3.159076	1.68963	4.50110
ann	1.5502066	1.233546	0.629116
prince harry	3.576400	1.728009	0.438358
...	....	....	....
....	....	....	....
....	....	....	....
george saunder	1.496208	0.512870	0.347217

```
term_ranks = np.argsort(np.argsort(-term_category_scores, axis
    =0), axis=0) + 1
metadata_descriptions = {term: '<br/>' + '<br/>'.join(
        '<b>%s</b> TextRank score rank: %s/%s' % (cat,
            term_ranks.loc[term, cat], corpus2.get_num_metadata
            ()))
        for cat in corpus2.get_categories())
    for term in corpus2.get_metadata()
}
```

Scattertext's dense-rank difference score makes it ideal for two-category contrastive plots because it sorts out phrases that are more associated with one category than another [Kessler 2017]. Thus, the chats in Figure 17 are constructed using standard dense-rank difference, a score that can be used in two-category contrastive plots and gives most category-associated phrases; the other uses the maximum category-specific score, which gives prominent phrases in each category regardless of the prominence in the other category (see Listing 4).

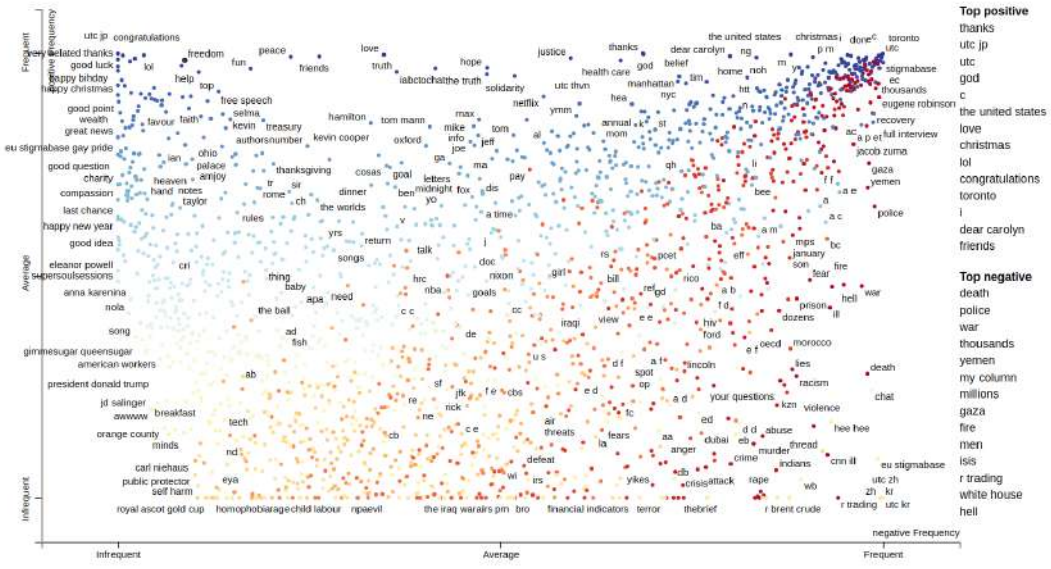


Figure 17: Phrases score extracted using PyTextRank algorithm [Mihalcea and Tarau 2004] for Twitter dataset.

Listing 4: Category-specific prominence

```
category_specific_prominence =
                        term_category_scores.apply
                        (lambda r: r.positive if r.positive >
                        r.negative else -r.negative, axis=1)}
```

In Figure 17, the two axes are the dense ranks of phrase centrality scores computed using PyTextRank [Mihalcea and Tarau 2004], while the terms scored are based on the sum of their category-specific eigenvector centrality as calculated by PyTextRank and the associate category is determined by the class's or category's average score that forms the highest. As shown in Figure 17, most of the associated terms in each class make some sense. For example, refer to (*thanks*) as the most identifiable preference term for authors (positive sentiment). On the other hand, *death* is the most negative preference topic users use in their tweets. Additionally, the results shown in Figure 17 also suggest that there is a certain degree of agreement between the centrality scores of terms across different categories.

Alternatively, we used the dense rank difference in the scores to color phrase-points and determine the top phrases to be displayed on the right-hand side of the chart. Instead of setting scores as category-specific prominence scores, we set *term_scorer=RankDifference()* to inject a way determining term scores into the scatter plot creation process as shown Figure 18.

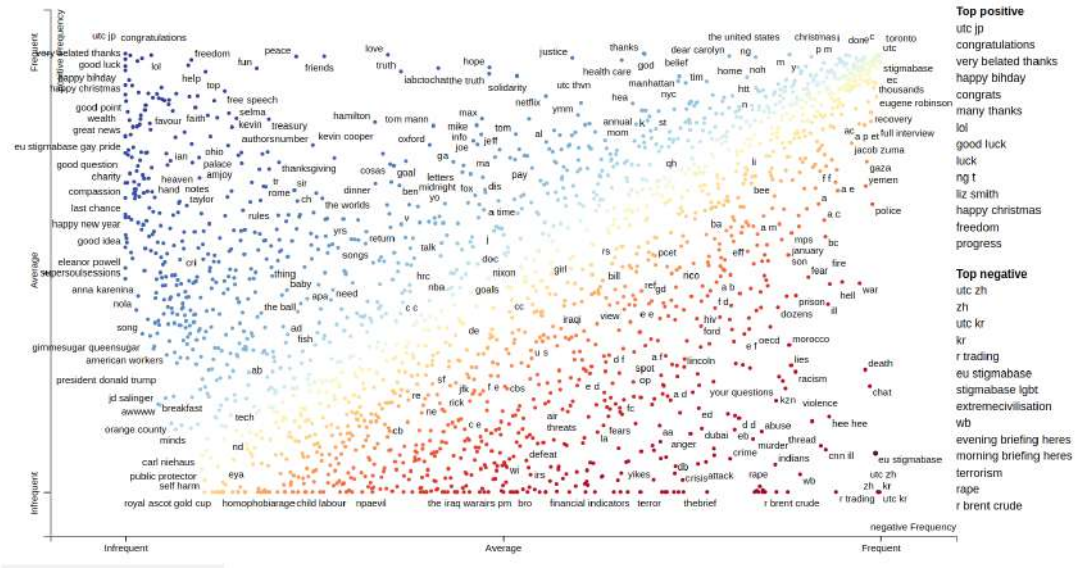


Figure 18: Phrases extracted using PyTextRank for Twitter dataset.

In Figure 18, the terms are scored based on the difference in the terms' dense ranks. These highlighted phrases are the most distinctive and characteristic of each category. For example, "*good luck*', was the most distinctive positive term used by most authors, and "*terrorism, rape*" was characteristic of a negative sentiment expressed by most authors in the texts. Furthermore, using regular expressions over POS tag sequences, we use phrase-based method described in [Handler *et al.* 2016] to identify frequent noun phrases in a given text based on unigram BoW. As shown in Figure 19, we select the 4000 most category-associated phrases from the Twitter dataset and plot them using densely ranked category-specific frequencies. The positions of terms are based on the dense ranks of their frequencies, and the scores, unlike Figure 18 that the terms' are dense ranks. Figure 20 and Figure 21 represent the phrases in the text as positive and negative, with appropriate sentiment scores. The negative term "*extreme civilization*", for example, has a sentiment score of -0.86182% with an acceptance rate of 20 words from 25,000

in the context. Similarly, with a 0.98588% acceptance rate, the positive term "*very belated thanks*" has 66 words from 25,000 in the context.

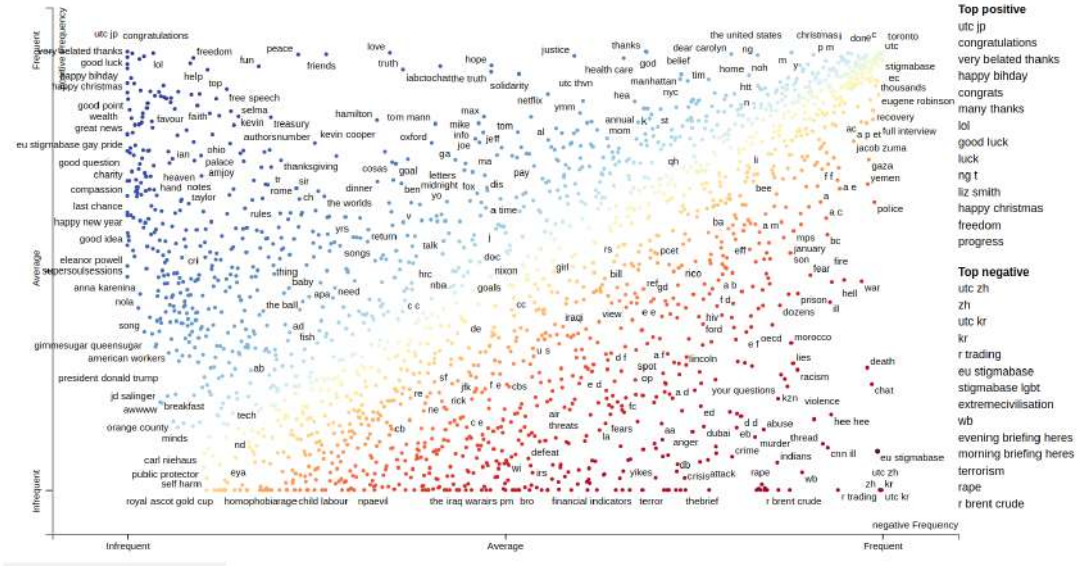


Figure 19: Phrase-machine [Handler *et al.* 2016] for Twitter dataset.

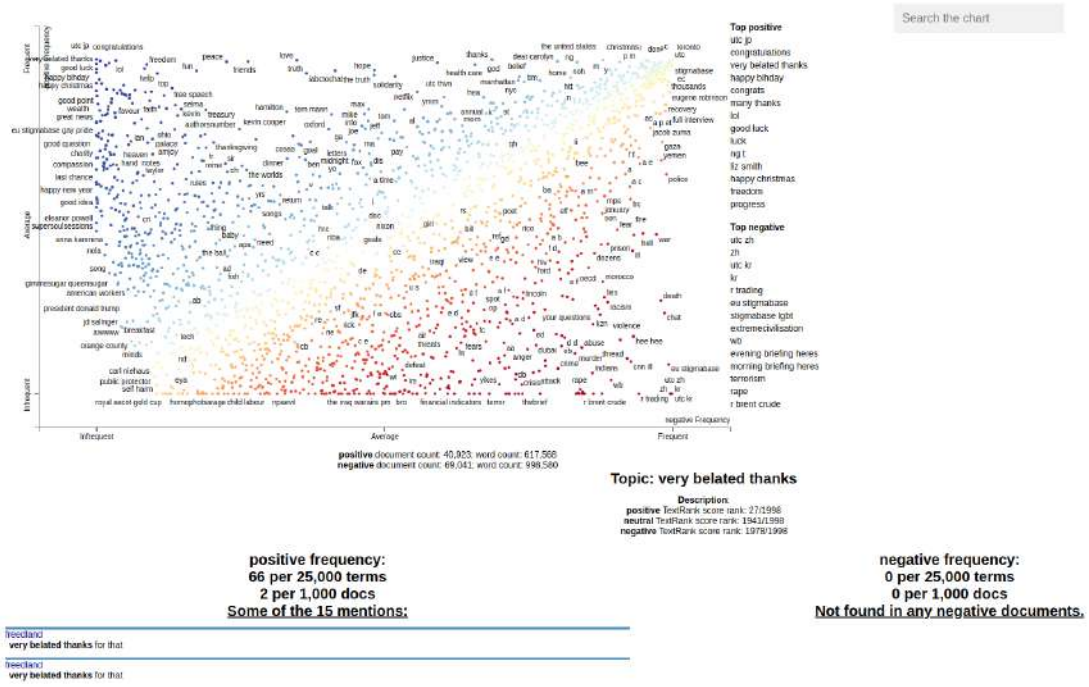


Figure 20: Example of Phrase-machine with positive sentiment terms in Twitter dataset.

3.3.5.2 *Visualizing topics signal*

The datasets in Section 3.2 are human textual conservations that cover a wide range of topics, so we use Listing 5 to visualize a wide range of preferred topics and their respective positions (e. g. categories). First, we create an instance of the corpus in order to extract topics and

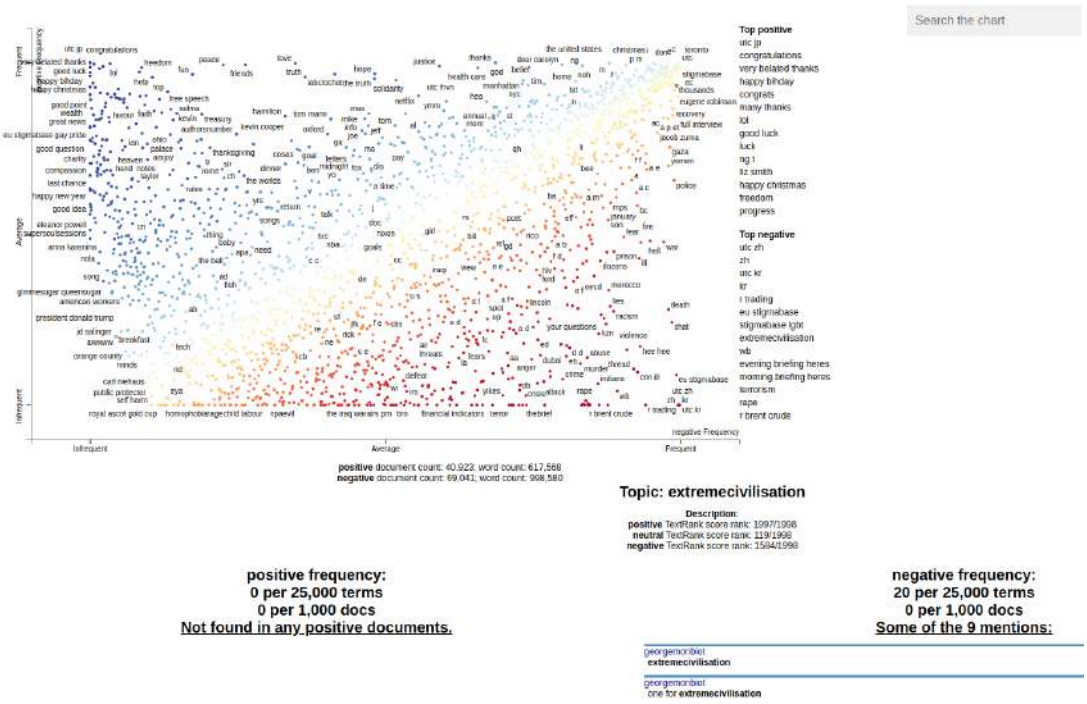


Figure 21: Example of Phrase-machine with negative sentiment terms in Twitter dataset.

categories instead of unigram and bigram features from the dataset. To do so, we use the function called "*FeatsOnlyFromEmpath*" from the scatter-text Python library [Kessler 2017]. The visualization is created by passing "*use_non_text_features*" as a true value argument into "*produce_scattertext_explorer*" and this instructs the Empath topics library [Fast *et al.* 2016] to look at the category rather than the terms to generate a snippet of statements that capture each author's choice of vocabulary or match the topic model (see Figure 22).

Figure 22 shows a stronger conveyed sentiment across both positively and negatively charged categories, as well as trends for exploratory discovery of interesting topics and categories in the Twitter dataset over sample text. Positive topics in the sample Twitter dataset use words like "affection" as top topics, with a sentiment score of about 0.76113%; they appear in 94% of the 25,000 terms and 27% of the 1,000 documents mentioned. Negative topics mentioned in the sample text, on the other hand, are associated with words such as textit"kill" or *"hate"* and the sentiment score is approximately -0.58813%, as mentioned in 42.19 documents per 25K words.

Listing 5: Topics and categories visualization

```
feat_builder = st.FeatsFromOnlyEmpath()
empath_corpus =
        st.CorpusFromParsedDocuments(tweet,
        category_col='vader_prediction',
        feats_from_spacy_doc=feat_builder,
        parsed_col='text')
        .build()
html = st.produce_scattertext_explorer
            (empath_corpus, category='positive',
            category_name='positive',
            not_category_name='negative',
            width_in_pixels=1000,
```

```
                metadata=tweet['target'],
                use_non_text_features=True,
                use_full_doc=True, topic_model_term_lists=
                    feat_builder
                .get_top_model_term_lists())
open(vis_path+"ConventionVisualizationEmpathTweet50.html",
                    'wb').write(html.encode('utf-8'))
```

3.3.5.3 *Query-based visualization*

Word representations have recently become a hot topic in NLP. While we have demonstrated and visualized how terms are related to one another based on their scores. In addition, we explore the possibility of visualizing a text-based categorical query (e. g. examining how text or document categories differ) from a Twitter dataset. We used spaCy libraries with GloVe word vectors [Pennington *et al.* 2014] trained on the Common Crawl corpus [Mehmood *et al.* 2017]. The cosine distance between vectors is used, with mean vectors used for phrases. The calculation of the most similar terms associated with each category is a simple heuristic. First, a set of terms closely associated with a particular author (e. g. category) is determined; secondly, these terms are ranked based on their similarity to the query, and the top-ranking terms are displayed using a scatterplot. A term is considered associated if the p-value is less than 0.05, which is a difference in the weighted log-odds ratios with an uninformative Dirichlet prior [Monroe *et al.* 2008]. Since we are scoring bigrams in addition to the unigrams scored by p-values, the size of the corpus must be larger to have high enough bigram counts for proper penalization. This function is based on the Dirichlet distribution's alpha parameter, which is a vector that is uniformly set to 0.01. Listing 6 shows the code that generated Figure 23.

Listing 6: Query-based category

```
from scattertext
import word_similarity_explorerhtml = word_similarity_explorer(
    convention_corpus,
            category='positive',
            category_name='positive',
            not_category_name='negative',
            minimum_term_frequency=5,
            pmi_threshold_coefficient=4,
            width_in_pixels=1000,
            metadata=tweet['target'],
            alpha=0.01,max_p_val=0.05,
                save_svg_button=True)
          open(vis_path+ "ConventionVisualizationJobsTwitter50.
              html",
          'wb').write(html.encode('utf-8'))
```

Figure 23 shows the words and phrases that are semantically similar to the word "incentives" in darker colored on a gray-to-purple scale, and general and category-specific related terms are listed to the right, which basically demonstrate topical category discriminators. This visual aid is helpful in understanding how the concept of incentives relates to other words, and provides a clear picture of how a given word relates to its broader semantic space.

3.4 CHAPTER SUMMARY

In Chapter 3, we have been able to carry out an analysis of widely used datasets to explore the use of exploratory statistical analysis and visualization to interpret a broad range of topics. This analysis revealed important insights about the language used by suspects in different contexts

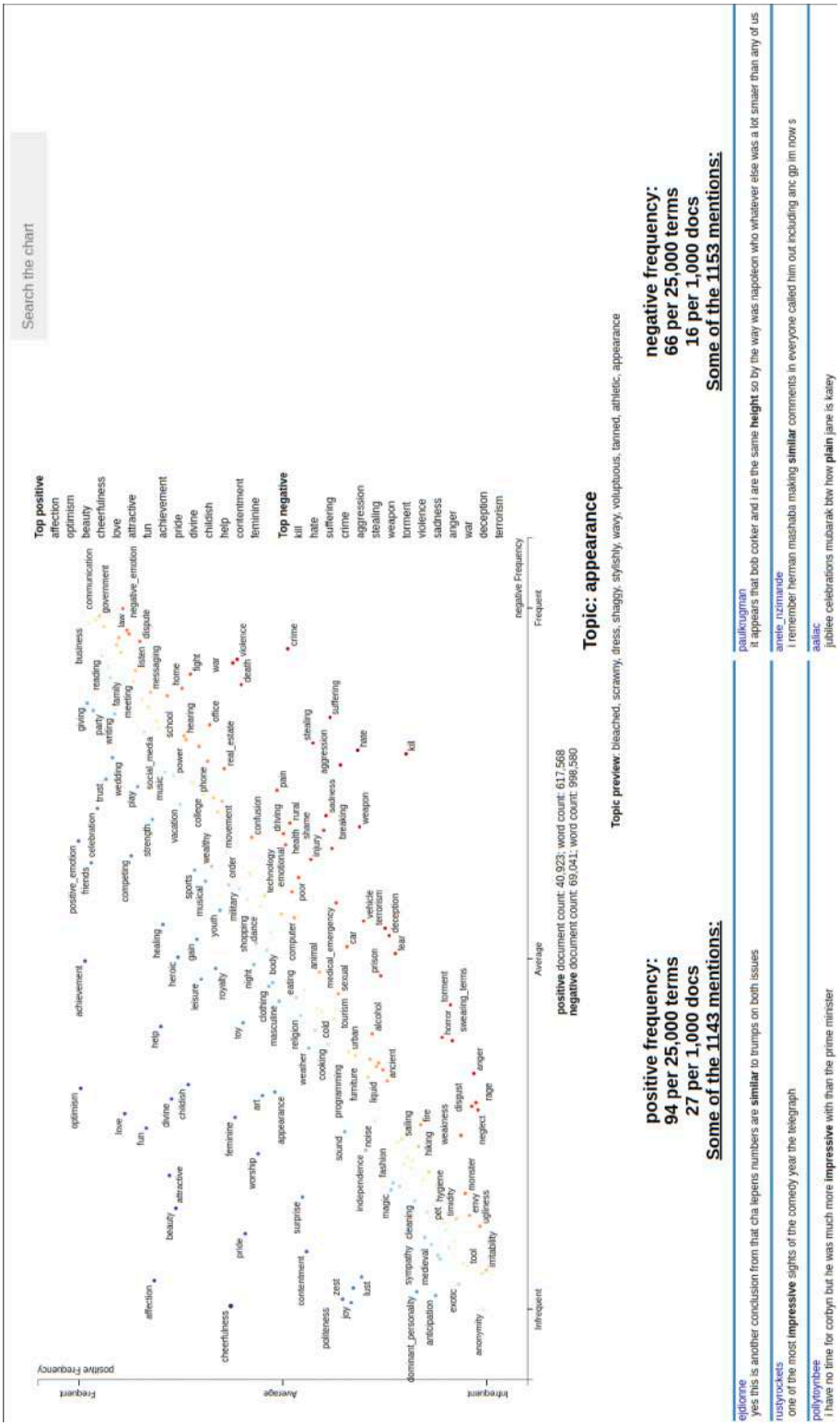


Figure 22: Example of topics and categories visualization over a sample text in Twitter dataset.

and allowed us to draw conclusions about their level of sophistication in expressing their thoughts. These insights were important because they helped us understand how suspects talk and gave law enforcement officials more information to think about when they were

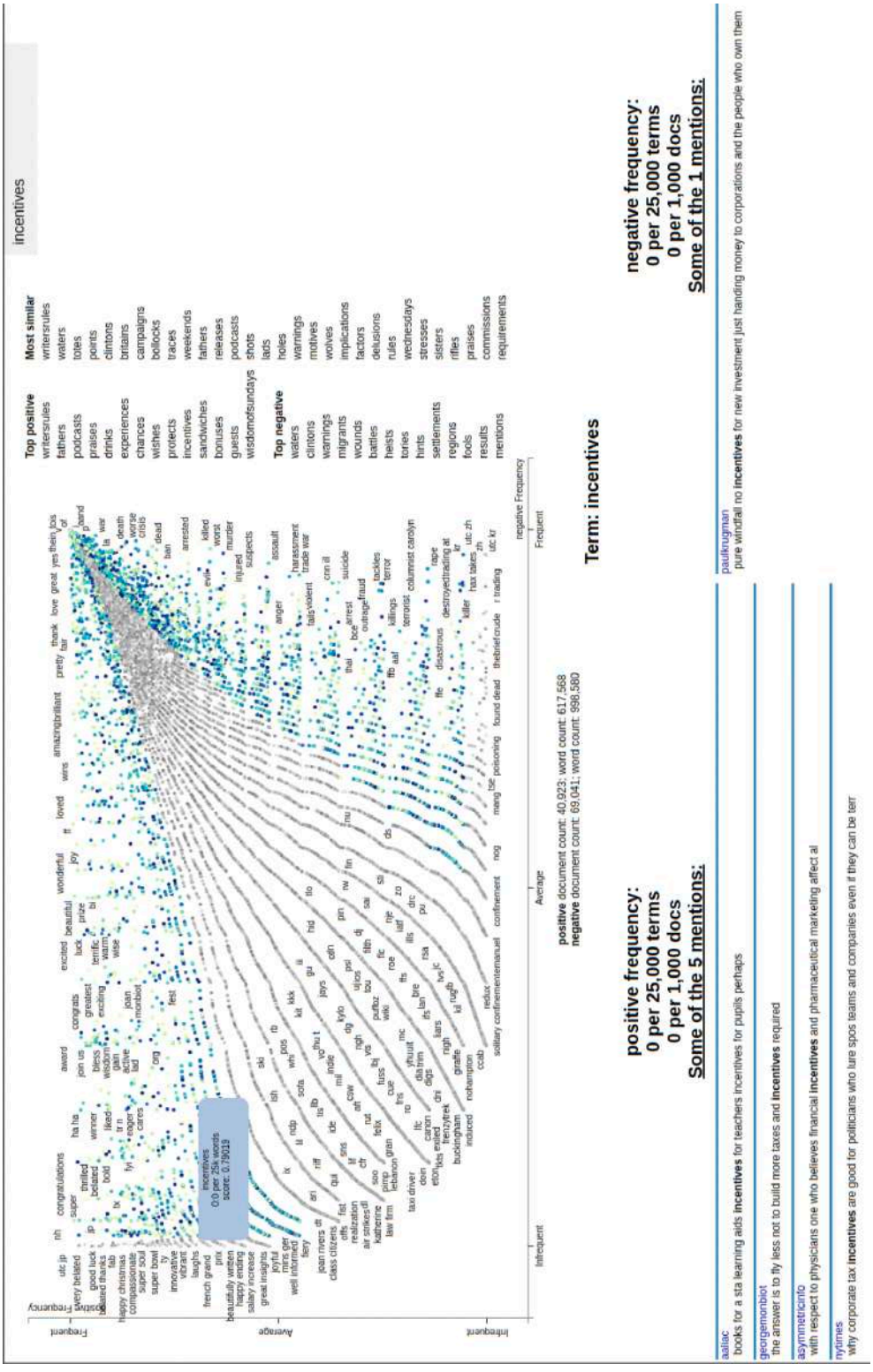


Figure 23: Visualizing query based on category different.

investigating. Furthermore, this type of analysis was beneficial because it provided more information about the context in which suspects were speaking and writing.

Part IV

DEEP REGULARIZED NEURAL NETWORKS FOR PAA TASKS

In Chapter 4, we define post-authorship attribution as the process of extracting identifying qualities from anonymous social media texts (e.g., Twitter, blogs, and online news stories) to identify the actual authors. In recent years, the field has grown in popularity due to its practical application in the justice system to identify the true authors of malicious texts and/or use the analysis results as proof of evidence in court without jeopardizing the individual's confidentiality. Most previous work has relied on discrete feature representations but fails to consider the underlying syntactic and semantic relationships between features [Mikros and Perifanos 2013; Joulin *et al.* 2017]. Additionally, the inconsistencies in datasets, mismatches, and evaluation methods make it difficult to measure the effectiveness of the methods [Tyo *et al.* 2022]. Moreover, the method also lacks the ability to offer interpretation, which is very important in many real-life situations, like forensic analysis or cybercrime investigation, which are needed to thoroughly explain the evidence in court.

Chapter 4 introduces a regularised deep neural network (RDNN) model as a key task for quantifying the writing style of the author from both social media text and contextual documents of literary text to improve generalisation in the domain of authorship attribution. The proposed regularised deep neural network (RDNN) model incorporates character-level features to extract stylometric elements. More specifically, the proposed RDNN model accepts a sequence of encoded characters (a sentence) as input to find the corresponding one-hot vector for each character to build character embedding and accumulate the results to form the input matrix. This model is fed into sequence-to-sequence models over convolutional distributed networks to learn a suitable writing style from online text snippets of social media and long sample texts of literary documents. The regularised deep neural network (RDNN) model surpassed state-of-the-art methods on four comparison datasets. It demonstrated its usefulness on three datasets, surpassing other methods while delivering equivalent results with the visualisation capability to interpret the features that contribute to the detection of the suspect author. This demonstrates the RDNN model's robustness and applicability across a variety of writing style categorization challenges.

4

REGULARIZED DEEP NEURAL NETWORK FOR POST AUTHORSHIP ATTRIBUTION

Post-authorship attribution is the process of identifying the author of an anonymous online text snippet (OTS) or lengthy text sample (LTS), such as literary writings, based on the author's distinctive writing style. Identifying the originator of these documents is critical to preventing cybercrime. Unfortunately, text length is limited on some devices, making the challenge more difficult. We introduce a regularized deep neural network (RDNN) to circumvent the intrinsic challenges of post-authorship attribution. It is based on a convolutional neural network, a bidirectional long-short-term memory encoder, and a distributed highway network. The neural network captures local and global lexical and stylistic features of the text, which are then sent to the bidirectional encoder to get a syntactic feature-vector representation. The feature vector was then supplied as input to the distributed high-speed networks for regularization to minimize the network generalisation error. The regularized feature vector was ultimately passed to the bidirectional decoder to learn the writing style of the author. The feature classification layer consists of a fully connected network and a SoftMax function to make the prediction. The RDNN method was tested against thirteen state-of-the-art methods using four benchmark experimental datasets to validate its performance. Experimental results have demonstrated the effectiveness of the method when compared to the existing state-of-the-art methods on three datasets while producing comparable results on one dataset.

4.1 INTRODUCTION

In recent years, the entire world has moved into a new modern period, with social media communication becoming increasingly important in our daily lives. Along with the above-stated benefits, various additional negative implications have emerged, one of which is the broad dissemination of false information for harmful reasons, with the user or individual publishing the content typically remaining anonymous and unable to be identified. Combating this issue, however, is tough, especially given the nature of OTS on social media platforms: a sea of tiny posts or messages and a large number of people. In Chapter 4, we address the problem of post-authorship attribution (a subset of authorship attribution) in OTS, which differs from what has been done with LTS by relying on a character-level approach for representing syntactic and semantic information in documents rather than discrete or handcrafted features that fail to generalise. We apply recent breakthroughs in deep neural networks to natural language processing and image recognition. By modelling OTS, a small text posted on social media, as unidimensional signals, we proposed a deep regularised neural network (RDNN) model to project the contentual information of OTS or LTS to model inherent linguistic knowledge in a natural language to understand and analyse the author's writing style from three perspectives: lexical, syntactic, and semantic, thereby combating this unwanted disinformation scenario.

In the rest of Chapter 4, we begin by discussing the methodology we implemented for authorship analysis tasks such as PAA or AA in Section 4.2. This method was based on regularised deep neural networks (RDNNs). Experimental results, including datasets, parameter settings, discussion of results in Section 4.3, and visualized results are explained in Section 4.3.4. We provide discussion in Section 4.4, and concluding remarks are provided in Section 4.5.

4.2 PROPOSED METHOD

The focus in this Chapter 4 is on the PAA subtask, which is generally hard to solve because of the inherent characteristics of the input text. The input texts are generally simple snippets

with short contents and are rife with obfuscating vocabularies such as idioms, onomatopoeias, homophones, phonemes, synonyms, acronyms, anaphora, and polysemy. However, it can be noted that none of the previous studies on PAA had focused on interpretability, which is a critical factor for real-life investigation scenarios. This current study further addresses the knowledge gap by adopting a method that utilizes the extracted features to visualize the results of the authorship attribution. Multiple studies have reported a higher authorship-attribution accuracy with several style markers such as lexical, syntactic, idiosyncratic, and structural markers [Rocha *et al.* 2016; Neal *et al.* 2017a]. However, the efficiencies of most of the past methods decreased with an increasing number of authors [Sarwar *et al.* 2018b]. They significantly depend on extensive feature engineering to reflect the content and style of an author [Ding *et al.* 2017; Stamatatos 2009b; Modupe *et al.* 2022]. They lacked transparency and interpretation capability [Wu *et al.* 2021a], which are essential for language-based pretrial investigation.

The regularized deep neural network (RDNN) was introduced in Chapter 4 to find possible ways of advancing the research on authorship-based tasks. The RDNN method promises to construct stylometry features by employing CNN, bidirectional long-short-memory (BLSTM), and distribution highway network (DHN). The purpose was to realize an improved PAA system with a high-precision visualization capability compared to the SOTA methods. The DHN guides the sentence structure in the LSTM cell to speed up the convergence rate of the BLSTM layers. Unlike the existing methods, the novel method can learn abnormal terms from combinations of characters, such as misspellings, emoticons, and stacks on two BLSTM layers. In addition, the BLSTM layer translates the exact text in reverse order to produce output, and the BLSTM layer is used to affirm the state of the past context of the text from a character-level CNN. Moreover, the DHN will allow the seamless flow of unimpeded information across the network before reaching the classification layer. The RDNN method was tested on four different datasets of text snippets to investigate its generalization capability across varying input modalities.

The core objective is to develop the regularized deep neural network (RDNN) method for improving the PAA system's accuracy. The proposed framework in Figure 24 is made up of three layers: a CNN character level layer, a distributed highway network (DHN) with a bi-directional long short-term memory (BLSTM) layer, and a feature classification layer. Incorporating the character-level feature with CNN, DHN, and BLSTM makes a PAA system do better at classification, convergence, and visualisation with high accuracy.

4.2.1 *CNN Character-Level Layer*

The CNN character-level layer environment includes three essential elements, such as character embedding, temporal convolution, and max over-time pooling. Character embedding is a numeric vector representation for a word constructed from the character n-grams. The PAA system receives a sequence of characters from a post as input and finds the corresponding one-hot vector representation for each character through the dictionary of m characters. The maximum length of a post is restricted to 512 characters, while the maximum number of sentences in a document is bounded at 280, owing to variations in the lengths of posts, and each sentence is filled to 512 characters in the preprocessing stage [Ruder *et al.* 2016a]. The English language training datasets were utilized in this study, so the dictionary contains 70 characters, including 26 English letters, 10 numeric characters, and 34 other characters, including white space, as follows:

abcdefghijklmnopqrstuvwxyz0123456789—,;.!? :′ "/\|_#$%&ˆ∗ ‘~ + − =<> ()[]{}

The CNN character-level layer of the proposed RDNN method is a slight variant of the deep CNN character-level [Kim *et al.* 2016]. In the CNN layer, each word in a post is replaced with

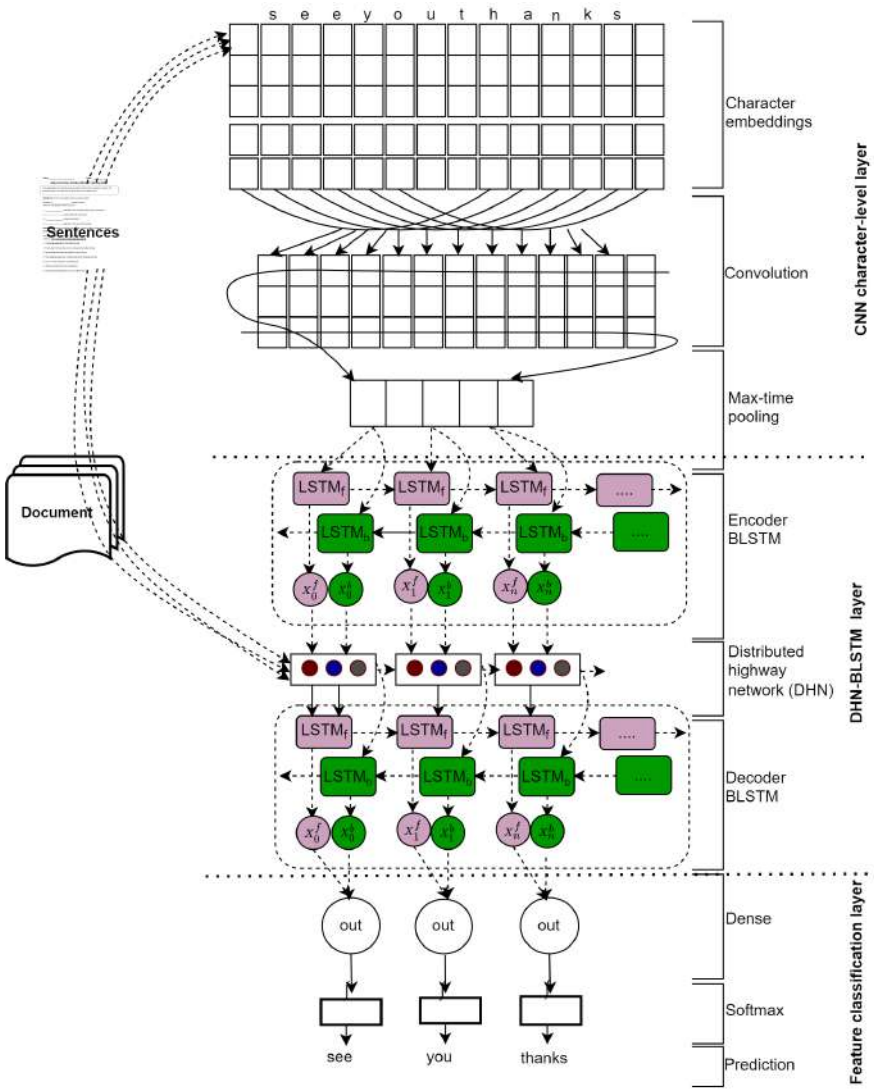


Figure 24: The architecture of the RDNN PAA system.

its one-hot vector representation. Hence, we denoted V to be the fixed vocabulary size of d being the dimensionality of character embedding and $X \in \mathbb{R}^{d \times |\mathbf{V}|}$ is the matrix of character embeddings. Suppose a word $w \in V$ of length l is made up of a sequence of characters $[c_1, c_2, \cdots, c_l]$, the character-level representation w is given by the matrix $C^w \in \mathbb{R}^{d \times l}$, where the j-th column corresponds to the character embedding c_j and c_j-th is the column of matrix X. The X^w matrix can be padded with zero for batch processing to ensure that the number of columns is equal to the maximum length of a sentence for all words V.

The temporal convolution exhibits longer memory than recurrent networks and is applied in this study to give better performance for character-sequence modeling. The convolution operations are performed between the input matrix and multiple filter kernels with stacking to form the input matrix. A new narrow convolution is used between X^w and a kernel $H \in \mathbb{R}^{d \times \omega}$ as a filter. Here ω is the width after adding a bias b and applying a non-linearity to obtain $f^w \in \mathbb{R}^{l-\omega-1}$ as a feature map. The i-th element of f^w is explicitly defined as follows.

$$f^w[i] = \tanh\left(\mathbf{X}^w[*, \mathrm{i} : \mathrm{i} + \omega - 1], H + b\right) \tag{6}$$

where $X^{\mathcal{W}}\left[*, \mathrm{i} : \mathrm{i} + \omega - 1\right]$ is the $i - \mathrm{to} - (i + \omega - 1) - \mathrm{th}$ column of X^w and $\boldsymbol{P,S} = \mathbf{Tr}\left(\mathbf{P}\mathbf{S}^{\mathrm{T}}\right)$ is the Frobenius inner product.

The max over time pooling operation is ultimately applied to get a fixed-dimensional representation of a word, which is output to the DHN-BLSTM layer. Equation 7 represent the max over time as the feature corresponding to the filter H to extract the highest value for a given filter. Therefore, each filter is essentially picking out a character n-gram that is the variation in the sentence such as lexical, syntactic, and topical structure with the size of the n-gram corresponding to the filter width.

$$y^w = \max_i f^w[\mathrm{i}] \tag{7}$$

4.2.2 *Distributed Highway Network with Bidirectional Long Short-Term Memory*

The distributed highway network (DHN) and bidirectional long short-term memory (BLSTM) are essential components in the second layer of the RDNN PAA system. Different from the existing models, DHN is added to the BLSTM to effectively regulate the flow of information and reduce the number of model parameters. The DHN features between the encoder BLSTM and decoder BLSTM guarantee the network information's validity. Moreover, they guide the sentence structure in the LSTM cell to speed up the convergence rate of the BLSTM layer. The encoder BLSTM was employed to find the state of the past context of a post from a character-level CNN. The decoder BLSTM will translate the same message in reverse order and produce output.

4.2.2.1 *Distributed Highway Network*

A distributed highway network (DHN) can optimize a deep-learning model with the core purpose of regulating the paths for information to follow across the different layers with a gate mechanism in LSTM [Srivastava *et al.* 2015]. In the feed-forward neural network (FFNN) consisting of L layers, and each layer use a non-linear transformation F with the weight matrix parameter W_F to generate the output $\boldsymbol{z}_i$ for the input x_i, and the tensor $\boldsymbol{z}$ can be represented as follows.

$$\boldsymbol{y} = F(x_i, W_F) \tag{8}$$

The DHN introduces two non-linear transforms, N and T in Equation 9 so that the outcome of $\mathbf{z}$ can be rewritten as follows:

$$\boldsymbol{y} = F(x, W_F) \cdot N(\mathrm{x}, \mathbf{W_N}) + x \cdot T(x, W_T) \tag{9}$$

where F is an affine transform followed by a nonlinear activation function, and N is the transform gate and $\mathbf{T}$ is the carry gate that express how much output can be produced by transforming and carrying the input. T is set to $N-1$ in Equation 9 to obtain Equation 10 as follows.

$$\boldsymbol{y} = F(x, W_F) \cdot N(x, W_N) + x \cdot (\mathbf{1} - N(x, W_N)) \tag{10}$$

The dimensionality of $x, y, F(x, W_F)$, and $N(x, W_N)$ must be the same to guarantee the validity in Equation 10. Consequently, based on the output of the transform gates, a highway layer can smoothly vary its behavior, and the outcome is used to adjust the vector from the CNN character-level layer.

4.2.2.2 *Bidirectional Long Short-Term Memory*

Long short-term memory (LSTM) captures the contextual information for solitary text sequences. However, the feature representation that is obtained from the CNN character level does not contain the sequence information. The bidirectional LSTM (BLSTM) is specialized in sequential modeling and can further extract the contextual information from the feature sequences obtained in the convolutional layer. Hence, we have used it to learn the representation of long-range dependencies based on the surrounding word distributions. The encoder

BLSTM was used to capture the lexical, and syntactic information from the CNN character level. The one-dimensional convolution involves a filter sliding over a sequence to account for features at different positions. Thereafter, the decoding layer operates on the encoded sentence representation with DHN. The decoder BLSTM will find the hidden structural representations for authorship attribution. The forward LSTM network computes the state $\overrightarrow{h}_t$ of the past left context of the sentence c_j, while the backward LSTM network reads the same sentence in the reverse order to produce $\overleftarrow{h}_t$ given the future right context. The outputs $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ outputs are concatenated to obtain the final output of the BLSTM network as follows.

$$h_t = [\overrightarrow{h}_t, \overleftarrow{h}_t] \tag{11}$$

The number of hidden layers for the input text is set to n, and the result of the decoder BLSTM network can be expressed as follows.

$$\mathbb{H} = [h_1, h_2, \cdots, h_m] \tag{12}$$

The parameter m is the length of the input text. The output of the LSTM layer $\mathbb{H} \in \mathbb{R}^{(m \times (2 \times n))}$ with each row of the matrix $\mathbb{H}$ representing a feature vector of the sentence generated by BLSTM. The hidden states of the time steps of BLSTM or the last layer of the time step hidden states of multilayer BLSTM are concatenated to obtain the feature representation of a post. The output of the highway network is used as input to the decoder BLSTM to learn the semantic representation of post sentences. The hidden document embedding of BLSTM is fed into a fully connected network for training and classification. The additional feature exploration was carried out to analyze the most important features for the overall performance of the RDNN method. The Shapley additive explanation (SHAP) [Lundberg and Lee 2017] and local interpretable model-agnostic explanations (LIME) [Ribeiro *et al.* 2016] for interpretability.

4.2.3 *Feature-Classification Layer*

The feature-classification layer consists of a dense layer or deeply connected layer from its preceding layer, SoftMax function, and prediction module. The feature-vector representation of a post is fed into a fully connected output layer that maps the weight vector of the decoder BLSTM to obtain the correct post classification as follows:

$$y(C_t|*) = \text{SoftMax}(\mathbb{H}_t, h_{t-1}) \tag{13}$$

Here, C_t is the feature vector for a post, the weight matrix h_{t-1} is a dictionary containing the sentence embeddings learned for each feature and $\mathbb{H}$ is a weight matrix that is learned to predict the appropriate class label.

The entire network is trained as an end-to-end model, such that a post written by the same author would result in small values as shown in Equation 13. Hidden feature vectors in Equation 12 must be made insensitive to topical variations between posts. The model is practically trained on randomly selected English language datasets without external resources. To control the generalisation error, extreme learning machine (ELM) [Ding *et al.* 2014; Huang *et al.* 2015a; Chen *et al.* 2017a] which were added to the random hiiden nodes one by one to achieve better classification results by incrementally updating the output weight. Then, use a dropout mechanism is applied to the regularization after the penultimate layer to learn the model parameters. The following cross-entropy loss is minimized during the training using the AdamW optimization algorithm Loshchilov and Hutter [2017]. AdamW generalises much better and able to compete with stochastic gradient descent (SGD) much faster when training compared to Adaptive optimizers like Adam [Kingma and Ba 2014].

$$CrossExt(p, q) = -\sum_{p} (d) \log(q(d)) \tag{14}$$

Here, p is the true distribution of one-hot vector representing characters in a post, and q is the output of the *SoftMax*. This corresponds to computing the negative log probability of the true authors.

4.3 EXPERIMENTAL RESULTS

The performance of the RDNN method is experimentally evaluated on large-scale English text datasets. The experiments were implemented using Ubuntu 18.04, Python 3.7 with Keras TensorFlow 1.15.1 application programming interface (API) on an Intel i7 4.0-GHz CPU computer with 64G DDR4 memory.

4.3.1 *Datasets*

As with the datasets discussed in Chapter 3, Section 3.2, the experimental datasets in this section use 10-fold cross-validation to divide the dataset into training, validation, and testing sets. This lets us evaluate the performance of the RDNN model compared with the previous SOTA approaches and assess its generalisation ability on unseen data.

4.3.2 *Parameter Settings*

The kernel sizes were 3, 4, 5, and 7 with the corresponding channel numbers of 64, 128, 256, and 512, respectively. The *ReLU* activation was used for a fair comparison. The dropout rate varied from 0.1 to 0.7, and the mini-batch size was set to 32. The number of hidden layers, m, in the DHN0-BLSTM layer was set to 512. The length of each convolutional layer was set to 128 for each kernel size. The SoftMax activation function was used in the feature-classification layer, and character embedding was initialized using the Glorot uniform initialization [Glorot and Bengio 2010a]. The entire model was trained for 60 epochs using the AdamW optimizer [Loshchilov and Hutter 2017] instead of Adam to yield a better training loss because it substantially generalizes better than Adam [Kingma and Ba 2014]. The learning rate was set to $(1e^{-4})$ and the dropout rate varied from 0.3 to 0.7. The number of layers in the DHN was set to 2. Table 5 reports all the hyperparameters in CNN-DHN-BLSTM architecture on all the different datasets. Here, d is the dimensionality of character embeddings, f, σ are non-linearity functions, ρ_s is the pooling size, δ_o is the dropout, l is the number of layers and m represent the number of hidden units.

Table 5: The setting of hyperparameters in our model on different datasets

Layers	Parameters	CCAT50	IMDB62	Blog50	Twitter50
CNN Character-level	d	30	70	70	70
	f	ReLU	ReLU	ReLU	ReLU
	δ_o	0.5	0.5	0.5	0.5
	ρ_s	2	2	2	2
BLSTM	m	128	128	256	512
	δ_o	0.3	0.3	0.3	0.6
FCL	l	2	2	2	2
	δ_o	0.7	0.5	0.5	0.5
	m	512	512	128	64
	σ	SoftMax	SoftMax	SoftMax	SoftMax

4.3.3 *Result*

The RDNN method is here compared to numerous SOTA methods for PAA with results presented quantitively and visually. The 5-fold cross-validation method with Keras tuner was used to find the best hyperparameters for the RDNN method. It utilizes 10% randomly split off as a stratified test set to minimize computational expensiveness that can occur with a higher number of folds. The best hyperparameters used to optimize the RDNN method are summarized in Table 5. A seed was used to reduce randomness by a further 10% for each training set. The quantitative performance metrics were the accuracy, F1-score, and loss curve calculated over 60 epochs with a 0.8/0.1/0.1 train/validation/test split. The visualization analysis was performed using the SHAP [Lundberg and Lee 2017] and LIME [Ribeiro *et al.* 2016] for interpretability.

4.3.3.1 *Quantitative Result*

The quantitative results based on accuracy (ACC) and F1-score (F1) performance measures are shown in Table 6.Table 6 presents experimental results from the literature for each of the models in comparison with the RDNN model with the same dataset. The evaluation metrics refer to the results reported with respect to the RDNN model so as to assess the effectiveness of the model. This helps to understand all the SOTA methods in relation to the RDNN model and helps researchers and practitioners identify potential improvements and advancements for future research efforts. Many of the results reported use different evaluation metrics; some use accuracy and F1-score, while others use either of the metrics to report their work. This is why in the Table 6, some are blank with some dataset and some are not. Also, most of the methods did not compute the measures across the four experimental benchmarked datasets. The RDNN method recorded good accuracy and F1-scores for three of the experimental datasets when compared to other SOTA methods. The proposed method concomitantly gave the highest accuracy and F1-score values of 93.20% and 92.20%, respectively, on the CCAT50 dataset. It recorded the highest F1-score of 76.69% and the second-best accuracy score of 63.40% on the Blogs50 dataset. In addition, it scored the highest accuracy of 86.48% with the second-best F1-score of 84.80% on the Twitter50 dataset. However, it came in second in accuracy (96.50%) and third in F1-score (95.75%) on the IMDB62 dataset when compared to the SOTA methods investigated. Its results are highly comparable with the top-performing methods, with the highest accuracy score of 97.72% [Ding *et al.* 2017], and F1-score of 97.90% [Murauer and Specht 2021] on the IMDB62 dataset, where accuracy and F1-score are represented as ACC and F1, respectively.

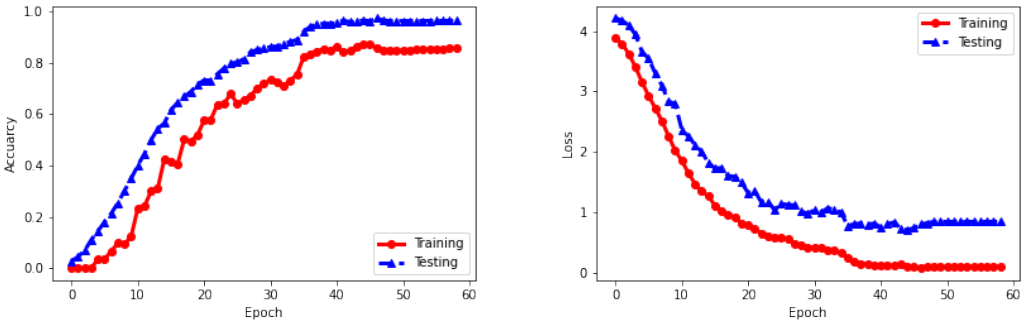
The effectiveness of the RDNN method is further presented graphically based on the accuracy and loss curves for visual validation purposes. Figure 25, Figure 25, Figure 26, Figure 27 and Figure 28 show the accuracy and loss curves for the CCAT50, IMDB62, Blogs50, and Twitter50 datasets, respectively. It can be observed from the figures that the proposed method generally has better performance in accuracy with faster convergence speed in the training and testing processes on all the datasets. However, complex interactions between different stylistic features with nonlinearities can be observed in the Twitter50 datasets. Although performance improvement is achieved with the RDNN method, the effect of certain features, such as hashtags, is highly indicative. This blurs the boundaries of the proposed method for understanding the writing styles of numerous authors. The Twitter50 dataset reveals a bias in the following ways: We did not preprocess the input post but converted it to lowercase to erode the character sensitivity effect of a possible discrepancy between upper- and lower-case letters. The RDNN intrinsically leverages the consequential effects of uniform resource locators, hashtags, and mentions of names in the posts to achieve high accuracy and an F1 score. The celebrities and power users used different vocabularies in their posts, as reflected in Figure 25 and Figure 27. This was, for instance, by frequently tweeting similar messages to enhance their brands in comparison to regular users or bloggers who were interested in news, as shown in Figure 26. However, this behaviour makes it hard for the

Table 6: Performance ACC and F1 of the RDNN method and SOTA methods for PAA sub-task.

Methods	CCAT50		IMDB62		Blog50		Twitter50	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Lexical & Topical [Ding *et al.* 2017]			97.20		41.60	41.60	82.50	82.50
PMSVM [Rocha *et al.* 2016]	80.45		90.51				64.80	
BertAA [Fabien *et al.* 2020b]			93.00		59.70	56.70		
SRNN [Jafariakinabad and Hua 2019]	90.58		94.10		61.19			
SCAP [Frantzeskou *et al.* 2007]				94.80		41.60		82.50
Imposters [Koppel *et al.* 2011]			49.90	76.90		22.60		52.50
SVM affix[Sapkota *et al.* 2015]	69.30							
SVM frequent [Plakias and Stamatatos 2008]	67.00	70.30	81.40					
Char & word *n*-grams [Schwartz *et al.* 2013b]							55.50	
Character *n*-grams [Schwartz *et al.* 2013b]	89.00	81.00	88.00	81.00	59.10	71.10	60.00	63.10
CNN word [Ruder *et al.* 2016a]	80.00	88.00	88.20	84.30	44.50	43.00	80.50	80.00
CNN character [Ruder *et al.* 2016a]	92.00	91.00	86.00	91.70	48.10	48.00	86.00	**86.00**
Syntax-CNN [Zhang *et al.* 2018b]	85.00	86.00	96.16	96.00	56.73	56.00	83.45	82.00
CNN *n*-gram [Shrestha *et al.* 2017]	76.50	77.00	91.21	92.00	53.09	53.00	65.85	66.00
Cont. n-gram[1] [Sari *et al.* 2017]	72.60	73.00	94.80					
DeepStyle [Hu *et al.* 2020a]	80.45	81.00	90.51	90.00	59.00	60.00	64.80	65.00
BRET [Murauer and Specht 2021]	77.00	66.20	**97.00**	**97.90**	66.20	66.00	61.40	61.00
DistilBERT [Murauer and Specht 2021]	71.32	66.50	97.60	97.00	60.40	60.00	26.80	27.00
RoBERTa [Murauer and Specht 2021]	75.60	65.90	**97.90**	97.00	41.70	41.00	65.90	66.00
LDAH-S & Topic Seroussi *et al.* [2011]	75.00	85.00	72.00	72.00	18.30	18.30	42.43	38.30
RDNN method	**93.20**	**92.20**	96.50	95.75	**63.40**	**76.69**	**86.48**	84.80

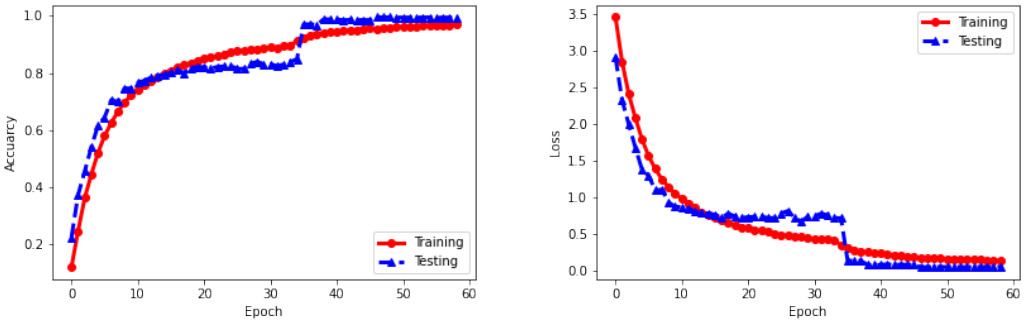
proposed method to consistently learn the writing style of an author. It would be prudent for future work to apply byte-pair encoding to investigate the stylistic differences between regular authors and celebrities and their impact on PAA methods [Vilar and Federico 2021]. Additionally, in Table 6, we only showed and highlighted the SOTA models that outperform the RDNN models proposed in Figure 24. However, the vast majority of SOTA approaches report utilizing either ACC or F1-score to demonstrate model performance; in this thesis, we exhibit both ACC and F1-score to highlight the success of our model.

With a set number of documents per author, as shown in Table 6, the performance of the RDNN models is very different when the dataset is split into training and testing sets, but neither the number of test documents nor the number of authors changes. The sampling of the documents was random, and all experiments were repeated five times to mitigate biases. It can be seen in Table 6 that the RDNN model is very superior in accuracy and F1-score for the Blog50 dataset, while it is less superior for F1-score with Twitter50 but raises sharply with accuracy. While the reason for this discrepancy remains unanswered through this experiment, it shows that most of the dataset and writing style of each author on social media platforms exhibit different behaviours when compared with the SOTA model and the RDNN model, perhaps as a result of restricted character space. Other issues could be the inconsistent text in the dataset, which might have made the cross-lingual large language models like BERT, RoBERTA, and others difficult to figure out the deep meaning in the piece of author writing as a sequence of words with a maximum of 512 tokens when training the models. While the input text from social media is very short and the RDNN model was scaled to a maximum



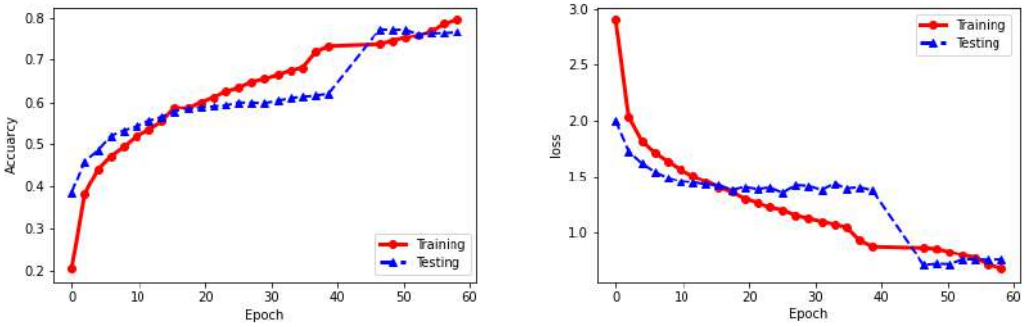
(a) Accuracy (b) Loss

Figure 25: Performance accuracy and loss curves over epoch on the CCAT50 datasets.



(a) Accuracy (b) Loss

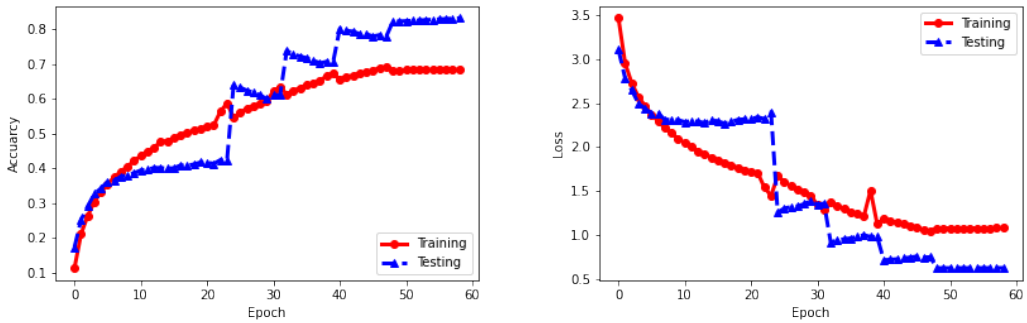
Figure 26: Performance accuracy and loss curves over epoch on the IMDB62 datasets.



(a) Accuracy (b) Loss

Figure 27: Performance accuracy and loss curves over epoch on the Blog50 datasets.

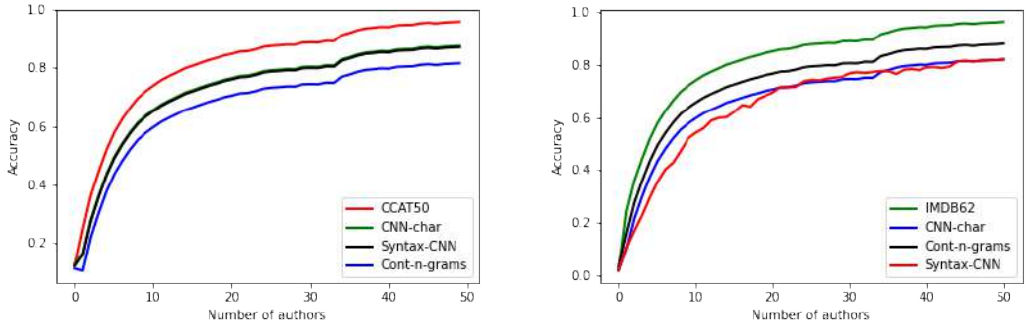
of 512, but extra information could have been added during training, which increased the computational constraint of the pre-trained model to reason semantically. Therefore, RDNN is implemented in the modelling of the innate author's writing style in natural language, analysing, understanding, and formulating linguistic intuition to uncover related behavior or patterns in other documents. Most of the results in Table 6 unambiguously demonstrate the struggle that pre-training large language models (PRT-LLM) have with distinct genres and topics such as social media text.



(a) Accuracy (b) Loss

Figure 28: Performance accuracy and loss curves over epoch on the Twitter50 datasets.

The performance of the RDNN method was further tested to determine how author number and post length increased with fewer iterations for all the experimental datasets. Figure 29 demonstrates how the RDNN method performs in comparison to the SOTA methods on CCAT50 and IMDB62 datasets. The RDNN method has superior accuracy in three experimental datasets and achieves comparable performance in the fourth experimental dataset when compared to the other SOTA methods investigated.



(a) CCAT50 (b) IMDB62

Figure 29: Performance accuracy on CCAT50 and IMDB62 dataset compared to SOTA methods with varying numbers of authors.

We also investigated the extent to which the RDNN methods could be scaled up for the Blogs50 and Twitter50 datasets, which had the lowest accuracy and F1-score (i.e., less superior), as shown in Table 6. Therefore, we did five independent tests on the dataset. The largest difference in the accuracy is less than 0.13%, which shows that the RDNN method is stable even though the accuracy isn't very high, as seen in Table 7 and Table 8. The results demonstrate that the RDNN method has superior accuracy and is far superior to the other PAA methods investigated in this Chapter 4.

Table 7: Performance variance through repeated runs on the Blog50 datasets

Metrics/Tries number	1	2	3	4	5
Accuracy (%)	63.39	63.35	63.44	63.38	63.46
Mean $\pm$ Standard Deviation	63.40 ± 0.045				
Variance	$\leqslant$ %0.112				

Table 8: Performance variance through repeated runs on the Twitter50 datasets.

Metrics/Tries number	1	2	3	4	5
Accuracy (%)	86.42	86.50	86.24	86.42	86.55
Mean $\pm$ Standard Deviation	86.48 $\pm$ 0.051				
Variance	$\leqslant$ %0.133				

4.3.4 *Result Visualization*

The visualization of results is useful for demonstrating how the feature vectors derived from the proposed RDNN system can assist skilled professionals in identifying an author's writing style for short and long documents, particularly brief text postings on social media. The purpose of the visualisation technique is to better understand how systems predict the results and find out who wrote text that has not been seen before from text that has been trained with the model. We also want to know what aspects of writing style, particularly features, are most important and contribute significantly to making the prediction. Model-specific and model-agnostic methods are the two primary groups of interpretables in any MLA or deep learning model [Molnar 2020]. Model-specific interpretation approaches are designed for specific types of models with the goal of understanding the internal structure and parameters in order to achieve appropriate interpretations. As a result, approaches such as Local Interpretable Model-Agnostic Explanation (LIME) and Shapley Additive Explanations (SHAP) have emerged to provide more detailed and individualised explanations and attributions than global interpretation techniques [Lundberg and Lee 2017; Ribeiro *et al.* 2016]. LIME is designed with the writing style representation extracted from the RDNN model with the LIME library, as shown in Listing 7. This helps us understand how the RDNN model behaves for the most relevant features in the experimental dataset.

```
1  LIMEexplainer=LimeTabularExplainer(
2        transform_2_lime_format(x_train,
3          ycat_name),
4      values,
5      mode="classification",
6      most_sign_features=x_train.columns.tolist(),
7      ycat_name=ycat_name,
8      cat_features=ycat_name.keys(),
9      discretize_continuous=TRUE,
10     random_state=42)
```

Listing 7: Pseudo-code for LIME explainer

The aforementioned parameters in Listing 7 are filled in by hand because the explanation lacks any defined values. First we call our now formatted dataset and set the mode to classification, then we give a list of all features in our dataset (line 3), and with ycat_name=ycat_name we specify which of the variables are categorical (line 4). *cat_features* as a categorical feature (line 5) lists the index of all features with a categorical (e.g., the lable or class), and discretize-continuous (line 6) is a mathematical function that simply helps to produce a better output by converting continuous attributes to nominal attributes. The final parameter, random state, brings consistency to the function; otherwise, it always picks a different number whenever we reload the function.

Executing the algorithm display in Listing 7 as the LIME output consisting of two parts: the prediction probabilities on the left side, the feature probabilities in the centre, and the feature-value table on the right. The prediction probabilities graph in Figure 30, Figure 31 and Figure 32 shows the RDNN model's decision (or predictions) from randomly selected instances of documents in the CCAT50, IMDB62, and Twitter50 datasets, meaning the outcome the

author predicts and the corresponding probability. The light green bar with the author name '*SamuelPerry*' represents the 41% probability as a style of writing or pattern from the instance that the RDNN model picked that originated from the author. The pink bar with the author name '*ThersePoletti*' represents the 21% probability, and so on. This basically represents a feature probability graph and gives insight into how much of a feature influence the authors have in the sample dataset. For this observation, the variable *comput* is the most influential factor and supports the prediction that it is part of the writing vocabulary that author '*SamuelPerry*' uses frequently. The second most important feature is '*technolog*', which contributes to the most choosing words that the author '*SamuelPerry*' is the originator of the instance of the given text. In this case, we display the top six characteristics in our output, but theoretically, all the features could be listed in order of their importance. As we can observe in Figure 30, Figure 31 and Figure 32, these represent the feature value table that categorizes the features by importance, including their weight as the actual value that this feature possesses in this observation with the feature names. For example, the first feature, '*comput*', shows 0.09 in this table, which represents 9% probability contribution of the features in the instance of the text in the CCAT50 dataset. It is coloured a light green, as it is influencing the model's decision towards a particular author named '*SamuelPerry*'. As demonstrated in Figure 30, Figure 31 and Figure 32 with different instances of the CCAT50, IMDB62, and Twitter50 datasets, LIME could not differentiate between the machine learning models used but displayed each of them in the same way.

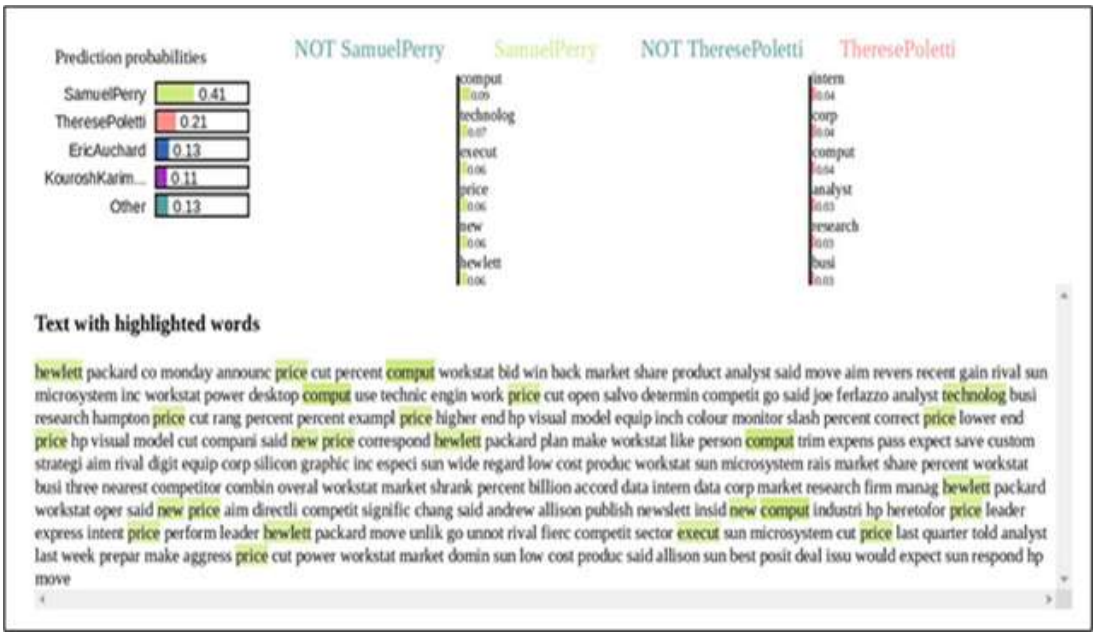


Figure 30: Explanation of individual prediction of the proposed method on the CCAT50 dataset using LIME.

It was observed for the Twitter50 dataset in Figure 32 that the RDNN method assigned the same prediction probability of 32% to the authors "*JacaNews*" and "*MaxBoot*" In addition, the method put the same weight on the function words "*american*" and "*britain*", which represent two different classes of authors, not *JacaNews* and *JacaNews*. The correct decision of the method shows that features such as "*america*" and "*wit*" are from *JaceNews* as an author, and words such as "*britain*" are more significant to *MaxBoot* as the author with a probability value. The SHAP [Ribeiro *et al.* 2016] over the feature vectors was utilized to obtain a good approximation of the SHAP values for the extracted features. The permutation was carried out by turning a particular word on and off. For example, if a word is not present in a particular post, turning it on could make a big difference to unveil the unknown writing style that could enhance the effective identification of an author. The non-existent feature in the document can receive a

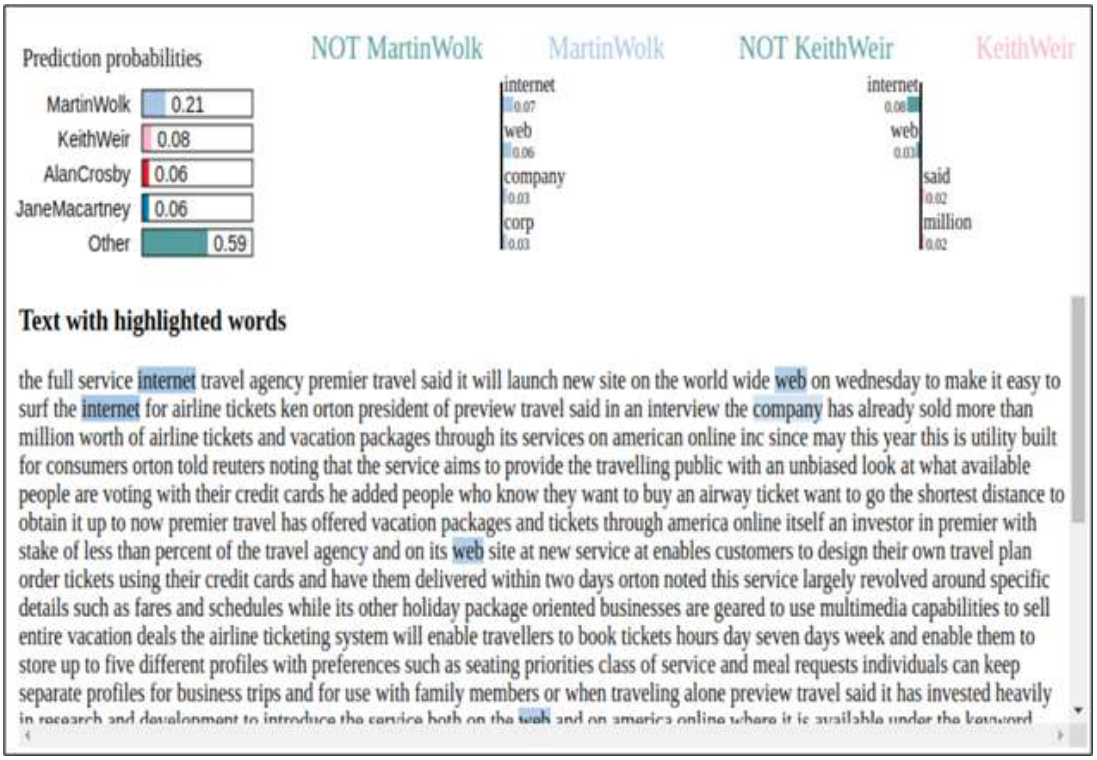


Figure 31: Explanation of individual prediction of the proposed method on the IMDB62 dataset using LIME.

high SHAP score. This was bypassed by applying permutations over the extracted features before the SoftMax.

Finally, the SHAP method is used to explain why certain traits were picked and to guess or group the real authorship of an online text snippet (OTS) in the Figure 24 for tasks that have to do with authorship, like PAA. To find the Shapley value, all potential sets of feature values with and without the *i*-th are compared. This displays how much each predictor or feature in the RDNN model contributes to the output features for each class label. To approximate the SAHP estimation for a single feature value, e. g.author writing style, as shown in Algorithmus 1. First, an instance of the dataset x, the feature index i, and the iteration step K are selected. Then, a random sample is selected from the data, and a random feature ordering is generated for each iterating step. By combining the values of x and r, two new features (or samples) emerged: instance x_{+1} as an interesting sample, as well as the number of possible coalitions exponentially increasing. Thus, [Štrumbelj and Kononenko 2014] proposed a solution to solve this kind of problem, but all values must be in the order after feature j is replaced by the actual feature values or author writing style learned from the RDNN model (or from sample r).

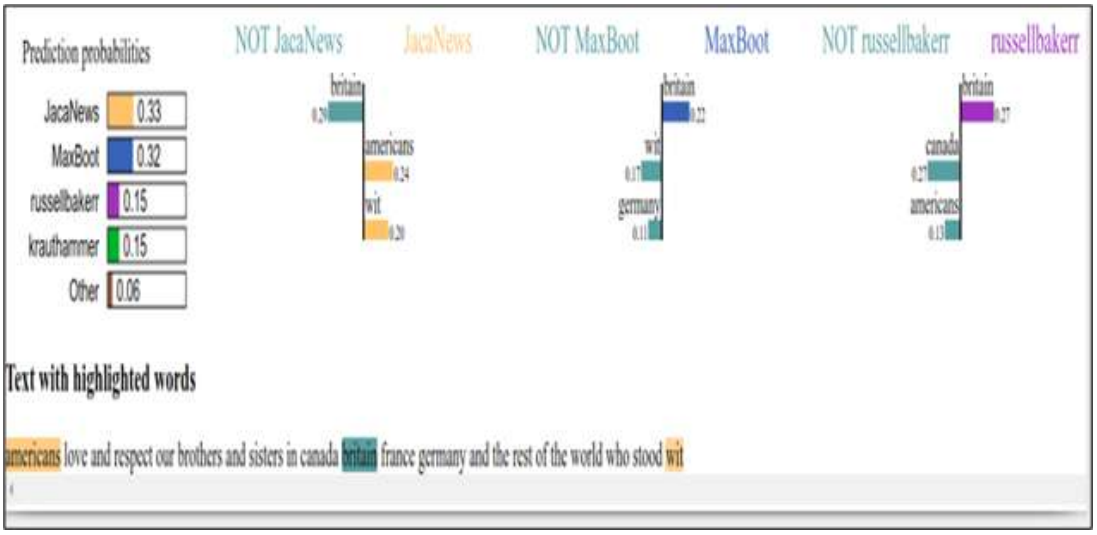


Figure 32: Explanation of individual prediction of the proposed method on the Twitter50 dataset using LIME.

Algorithmus 1 : Explanability of RDNN model with SHAP

Data : RDNN model f, instance x, feature index i, steps k and data matrix $\mathcal{D}$
Result : SHAP value for the value of the i^{th} feature $\mathcal{S}$

for $k \leftarrow 1$ *to* K **do**
  choose instance r in $\mathcal{D}$
  order instance $x : x_0 = (x_1, \cdots, x_i, \cdots, x_p)$
  order instance $z : r_0 = (r_1, \cdots, r_i, \cdots, z_p)$
  while $f \neq 0$ **do**
    if $f > 1$ **then**
      $i : x_{+i} = (x_1, \cdots, x_{i-1}, x_j, r_{i+1}, \cdots, r_p)$
    else
      if $f < 1$ **then**
        $i : x_{-i} = (x_1, \cdots, x_{i-1}, x_j, r_{i+1}, \cdots, r_p)$
      compute marginal contribution:
      $c_i^k = f(x_{+i}) - f(x_{-i})$
      calculate the SHAP value as the average
      $c_i(x) = \frac{1}{K} \sum_{i=1}^{K} c_i^k$
      $\mathcal{S} \leftarrow c_i(x)$
return $\mathcal{S}$

After teaching the RDNN model and getting the prediction or classification results, we used SHAP with the Algorithmus 1 to figure out what the classification decision meant. For example, we looked at the selection to explain the best feature or the author's writing style (e.g., feature values) that was used to find a certain author. Figure 33 and Figure 34 demonstrate how the training set's chosen writing style affected the first random sample on the CCAT50 and IMDB62 datasets as well as the SHAP scores obtained using the RDNN method. As shown in the graph, the base value represents the average of the RDNN model estimation results based on the training set. If the output of the model is to the right of the base values, then it represents that the values obtained are higher than the base value. Red arrows indicate the degree of influence of input variables that increase the occurrence of certain writing styles or feature values to predict a particular author. Blue arrows indicate the influence of input variables that suppress the occurrence of a particular author based on their

writing style. These arrows provide insights into the factors that contribute to the occurrence or suppression of specific writing styles or feature values when predicting a particular author. By analyzing the direction and magnitude of these arrows, we can better understand the impact of different input variables on the model's predictions.

Figure 33 and Figure 34 show local explanations where the probability of the output is 86% for CCAT50 and 75% for IMDB62 datasets, along with the author's writing styles or feature values shown below ('have', 'own', 'beginning', $\cdots$, 'getting'$\cdots$,'some') and so on. Figure 33 shows the case with the highest prediction with the indicator that the potential feature contribution is the sequence of words in blue. In particular, the sequence order of the words reflects the significant contribution of the vocabulary in predicting the author as illustrated in Algorithmus 1.

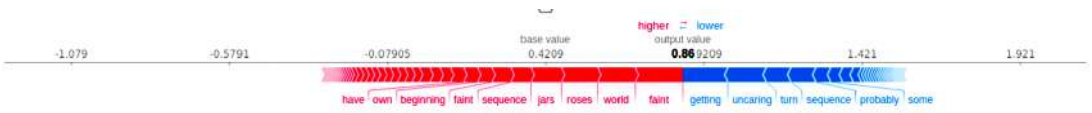


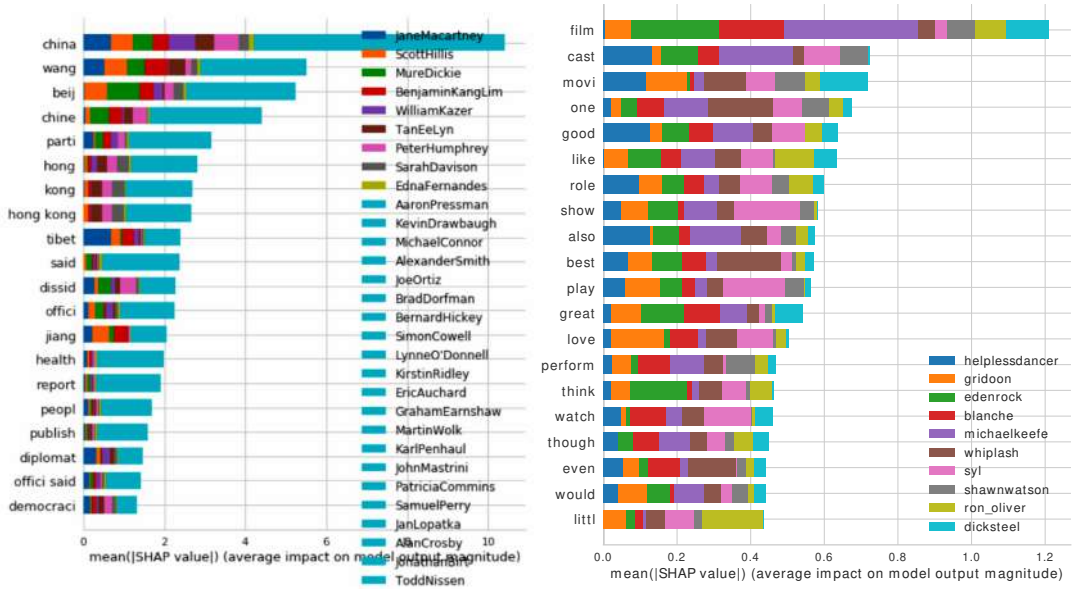
Figure 33: Explanation of evaluation results based on the SHAP tree explainer for CCAT50 dataset.

Figure 34: Explanation of evaluation results based on the SHAP tree explainer for IMBDB62 dataset.

Figure 35 show the SHAP summary plots for the global explanation of the model. The plots combine the features with important effects based on the sample texts from CCAT50 and IMDB62 datasets. Shapely value for a feature and particular sample is represented by a point on a summary plot. Features are on the y-axis and shapely values are on the x-axis. The colors are used to represent low or high values. Features are arranged according to their importance, and the peak feature in the summary plot is the most important, whereas the valley feature is the least important.

4.4 DISCUSSION

The best performance result for each SOTA method is reported here, as presented in the relevant articles. This is because the source codes of the comparative methods are unavailable. The proposed method outperforms most of the investigated SOTA methods, including the deep-learning methods, with an accuracy margin of 2.62%, 0.34%, and 3.98% on the CCAT50, IMDB62, and Twitter50 datasets, respectively. However, it is less superior to the top-performing method by 2.3% on the Blogs50 dataset. The accuracy of the method with the SVM frequent 3 n-grams [Plakias and Stamatatos 2008] did not perform well with the lexical, character, and syntactic features on the CCAT50 dataset when compared to that reported by [Sapkota *et al.* 2015]. In contrast, the SCAP, LDAH-S, and CNN-word methods perform well in domains such as the CCAT50 and IMDB62 datasets where topical information is discriminatory. However, they achieved a comparatively worse accuracy for short messages such as the Blogs50 and Twitter50 datasets. It can be suspected that SCAP stores discrete n-grams for each author, which makes it simple to distinguish a topic preference for each author. However, it is unsatisfactory to capture the writing style of different authors or identify several posts written by an author. If such a correlation exists, the low accuracy reached by



(a) CCAT50 (b) IMDB62

Figure 35: SHAP summary plots for CCAT50 and IMDB62 dataset

the Syntax-CNN [Zhang *et al.* 2018b], SRNN [Jafariakinabad and Hua 2019], and DeepStyle [Hu *et al.* 2020a] proves that the models carry less topical information. Moreover, the stylistic representation loses cohesiveness when the number of authors is very large.

The RDNN method is still superior in capturing better stylistic information such as words, topics, punctuation, or terms. In terms of the average F1-score, the RDNN method outperforms the comparative SOTA methods on CCAT50, Blogs50, and Twitter50 datasets by a margin of 21.9%, 19.9%, and 2.0%, respectively. However, the BertAA model is highly competitive on the IMDB62 dataset by roughly 2.15%. The word-based methods such as SVM, BertAA, CNN-char, and CNN-word perform well on CCAT50 and IMDB62 datasets, in which topical information is unprejudiced. However, our method achieves a comparative performance boost for Blogs50 and Twitter50 datasets. Even the shallow neural network of FastText embedding [Sari *et al.* 2017] does not beat the SVM histogram [Frantzeskou *et al.* 2007] with similar feature sets. Nevertheless, our PAA method shows superior encoding ability for the task of authorship attribution on both CCAT50 and IMDB62 datasets. The fine-tuned BERT, BertAA, and continuous$n$-gram representation model with pre-embedding vectors are sensitive to topical writing divergence between authors, which boosts performance on the IMDB62 dataset. However, they are less effective on the Twitter50 dataset than the CNN-char, where hashtags or emoticons are the most distinctive features. Moreover, our method outperforms the SVM model with frequent 3 n-grams, LDAH-S, Imposters, CNN variants, and other deep neural network learning methods on three of the four datasets. This is in terms of accuracy and F1-score performance measures, except for the Twitter50 dataset, where CNN-char outperforms our PAA method in F1-score by 2% for 50 authors. The experimental results have demonstrated the aptitude of the RDNN method to handle the class-variation problem inherent in real-world text-communication applications [Stamatatos 2009b].

The visualization plots of Figure 33 and Figure 34 show the prediction for just one example of a testing dataset using the RDNN method. The SHAP values from the "based" value represent the features that are predicted from the RDNN method to reflect the writing style of an author. The analysis of how SHAP values are attributed to the writing style of an author is provided in Figure 33 and Figure 34. In this case, all features are continuous and vertically sorted by the average impact on the predictions. It can be observed that "china" and "film" have the

maximum impact on the prediction with high mean values on CCAT50 and IMDB62 datasets. Similarly, high mean values recorded by “wang”, “beji”, “cast” and “movi” on CCAT50 and IMDB62 datasets are significant and related to the attributions of a writing style of an author in concordance with Figure 33 and Figure 34.

4.5 CHAPTER SUMMARY

Post-authorship attribution is still an open research problem because of different kinds of intrinsic ambiguities in text snippets. In this Chapter 4, a character-level authorship-attribution method is proposed based on a convolutional neural network, a distributed highway network, and bidirectional long short-term memory. The proposed method can efficiently extract lexical, syntactic, and structural representations from a given post to identify the authorship of questionable texts. The proposed method has been extensively evaluated and compared against some of the state-of-the-art methods to report new results across the standard datasets. Some of the state-of-the-art methods performed comparably well in certain cases because they were designed with a simple structure, but they cannot be used as evidence in courts of law. At the same time, the proposed method eliminates intrinsic fluctuations and produces high-quality interpretations. In contrast, we have shown how results obtained from the proposed deep-learning method can be interpreted using an explanation technique. Furthermore, a framework to identify the most influential stylistic features through a local explanation is provided that can be interpreted intuitively and is a useful tool for forensic investigators. The current study identifies some minor impediments that should be addressed in future research to significantly improve the performance of the proposed method. The effect of certain features, such as hashtags, blurs the boundaries of the proposed method for understanding the writing styles of numerous authors. The possible user behaviour of repeated tweeting of similar messages makes it hard for the proposed method to consistently learn the writing style of an author.

Part V

APPLICATION OF SUB-WORD EMBEDDING FOR PAA TASKS

5

AUTHORSHIP IDENTIFICATION BY INTEGRATING BIDIRECTIONAL LONG SHORT-TERM MEMORY WITH SUBWORD EMBEDDING

In Chapter 4, we talked about using a regularized deep neural network (RDNN) to learn an author's writing style by feeding the input data from the character level into a convolutional neural network, a bidirectional long short-term memory encoder, and a distributed highway network. Also, the RDNN model is a valuable tool for forensic investigators because it can find the most influential stylistic features through a local explanation for intuitive interpretation. Thus, the proposed RDNN model's convergence rate significantly speeds up the PAA system's performance. However, we can learn a writing style of an author from character levels and some parts of words too. However, social media (e. g.Twitter) and discussion forums (e. g. blogs) are popular because they generate large amounts of text data every million seconds. In this use case, datasets or documents with many informal or slang words that are not part of the standard vocabulary would have a false-negative feature vector. Also, these slang words are often used in short text messages from social networks. As a result, the RDNN model disregards how words naturally fit together in a text (for example, the sequential relationship of words that is inherent in the texts).

In Chapter 5, we propose a bidirectional long short-term memory (BLSTM) with a 2D convolutional neural network (CNN) over a two-dimensional pooling operation to capture sequential writing styles and introduce Gaussian noise to preserve the sequential context of the text to improve authorship-related tasks. We extensively evaluate the model that leverages subword embedding and compare it against state-of-the-art methods for an extensive range of authors. Our proposed methods improve 1.07%, and 0.96% on CCAT50 and Twitter, respectively, and produce comparable results on the remaining one.

5.1 INTRODUCTION

Authorship attribution (AA) is the process of identifying the author of anonymous texts and obtaining information about them. This field has attracted attention due to its relevance to a wide range of applications, including digital forensics investigation (e.g., identifying the author of anonymous texts or phishing emails [Iqbal *et al.* 2008b; 2013; Modupe *et al.* 2012b; 2014], plagiarism detection [Foltýnek *et al.* 2019; Layton *et al.* 2015], or historical literature [Neal *et al.* 2017a; Fabien *et al.* 2020b].

Over the years, AA has been studied on a long-text document involving text samples containing thousands of words [Ramnial *et al.* 2016a; Wanner and others 2017; Sari *et al.* 2017] as well as short text, such as tweets with text samples of a few words [Schwartz *et al.* 2013b; Azarbonyad *et al.* 2015; Rocha *et al.* 2016; Iqbal *et al.* 2013; Rocha *et al.* 2016; Seroussi *et al.* 2014; Koppel *et al.* 2011]. The extraction of important information characterizing the writer's identity is among the most challenging tasks in AA. Most of the methods proposed rely heavily on a large number of stylometric features to represent both the author's content and writing style. The hypothesis behind this feature extraction is that each author has a unique style of writing such as punctuation, the average size of words, the proportion of digits and capital letters, hapax-legomena, part-of-speech (POS), vocabulary richness, frequency of punctuation marks, functional words (FWS), and character/word n-gram [Madigan *et al.* 2005b; Aborisade and Anwar 2018; Seroussi *et al.* 2014; Fabien *et al.* 2020b]. However, they can hardly capture the syntactic and semantic information in the texts, which provides a deeper insight into the author's writing style. Additionally, several traditional methods of text

classification, such as topic methods, which are generally Latent Dirichlet Allocation (LDA) based on bag-of-words (BOW) [Stamatatos 2009b] or a simple statistics of some ordered word combinations [Cambria and White 2014; Poria *et al.* 2016] are prone to serious issues with optimisation, noise sensitivity, and instability to learn complex relationships. Some of the methods work well to find the author of a text but tend to malfunction when the text is short by not taking into account word order [Muttenthaler *et al.* 2019; Luyckx and Daelemans 2011; Sarwar *et al.* 2018b; Alonso-Fernandez *et al.* 2021].

Compared to topic modelling, deep learning neural network models have been proposed to learn more effective vector representations of words, e.g., the pre-trained word vectors, which are mapped into a vector space with semantically similar linguistic features to identify the author's writing style, have similar vector representations [Cambria *et al.* 2015; Zhou *et al.* 2015; Ruder *et al.* 2016a; Ding *et al.* 2017; Sari *et al.* 2017]. Also, these methods usually work as well as traditional ones, and they do not require any knowledge of the author's knowledge of syntax or semantics in the examples they use [kalchbrenner2014convolutional, kim2016character, lai2015recurrent, gupta2017study, zhange2018syntax, Jafariakinabad09723, wu2021exploring]. However, these models generally converge slowly. In addition, the embedding of rare words is poorly estimated, and the result would likely have negative effects on the representation of words surrounding the author's writing and the performance of the model.

In Chapter 5, we introduce bidirectional long short-term memory (BLSTM) with 2D convolution and a 2D max-pooling operation that employs byte-pair encoding (BPE) to transform given examples of texts into subword embedding. First, we use BLSTM layers to capture inherent semantic features on both the time-step dimension and the subword feature dimension. And then feed the feature vectors into a 2D CNN and 2D max pooling to obtain more local syntactical information to represent the input text for AA tasks. The features are converted to numeric representation using the BPE algorithm, which brings balance between character and word-level hybrid representation to build a learning language model for AA tasks. The main contributions in Chapter 5 are summarized as follows:

1. Usage of sub-word based representation supported on a byte-pair-encoding (BPE) algorithm to deal with imaginary word problems and splits the text into characters for the AA tasks. The approach is interesting to find n-grams based on the input text.

2. We proposed the use of Gaussian noise in the layers to reduces the number of parameters in the training and improve the performance for the AA tasks.

3. The experimental comparison of the proposed method with other existing prominent methods to demonstrate its performance.

The other part of Chapter 5 is organized as follows. In Section 5.1 is an introductory messages, and Section 5.2 we provides an overview of related work. In Section 5.3, we describe the structure of our proposed BLSTM-CNN max-pooling model for AA tasks in details. In Section 5.4, we described the details about the dataset, hyper-parameters setting and the experimental results. In Section 5.5, we show a brief discussion based on the outcome of the proposed model, and in Section 5.6, we concludes.

5.2 RELATED WORK

Traditionally, the AA task largely relies on extracting stylometric features related to content or style to ascertain the writer of the text. Most of the existing methods based on stylometric features aim to capture writing patterns in different linguistic categories, e.g., lexical, syntactic, structure, and semantics. Lexical features capture an individual's characters and words to describe vocabulary richness and choice of symbols or words. At the character level, features include the number or frequency of different characters. The word-level feature combines the total number of words, average word length, a portion of short/long words, most common

words, and the number of unique words [Argamon and Levitan 2005; Juola 2007; Stamatatos 2009b]. Another practical set of characteristics combined with lexical n-grams. They describe sequences of n elements next to one another. The elements can be different, for example, a sequence of characters, words, symbols, and syllables. However, as the dimensionality of the n-gram vector spaces grows with n, the character or word n-gram features capture too much content-specific rather than related stylistic information [Alonso-Fernandez *et al.* 2021]. For instance, Plakias and Stamatatos [2008] used tensors of the second order with 2500 most frequent 3-grams to represent stylistic components for a given text and Muttenthaler *et al.* [2019] show the influence of punctuation marks with an n-gram model for AA tasks while masking punctuation marks with an asterisk $(*)$ symbol. Markov *et al.* [2017] allege that digits and named entities are other critical features for selecting writing styles.

Sapkota *et al.* [2015] connect the potential of using character-level n-gram features to the high priority of subword features (e.g. suffixes and prefixes) in authorship-related tasks. Zhao *et al.* [2019] relax the constraint of n-gram features and analyze the co-occurrence of word pairs instead. As text representations created from the n-gram model tend to be high-dimensional and sparse, and Niu *et al.* [2017] use principal component analysis (PCA) to decrease them into low-dimensional vectors.

Seroussi *et al.* [2011]; Zhou *et al.* [2018] use a collaborative method to find the style similarity based on topic models. Structural features capture the organization of paragraphs and sentences. They include the number of sentences, paragraphs, lines, punctuation, and the average length of sentences and paragraphs. With structure features, elements such as greetings and signatures in a text can be analyzed. Similarly, Koppel *et al.* [2011] used lexical and structural features with multiple randomized characteristics to unveil the writing similarity between two authors by ignoring the order of words, syntax, or meaning in the text. However, in recent time many researchers have turned to neural language models such as the skip-gram model [Mikolov *et al.* 2013a;b; Pennington *et al.* 2014] based on the distributed representation of the words to learn the distribution of the writing style [Ding *et al.* 2017; Gómez-Adorno *et al.* 2018b; Posadas-Durán *et al.* 2017].

On the other hand, syntactic features characterize the use of punctuation and function words (FWS), which help define the relationship between elements in a sentence. It also includes POS tagging by categorizing a word as either a verb, noun, pronoun, or adjective (e.g., according to its function) Bevendorff *et al.* [2019] develop character tri-gram vectors for the documents and evaluate the variations between each couple of documents as features using seven distance measures. Bagnall [2015] employs a recurrent neural network (RNN) model on the character level to verify authorship and obtain a higher accuracy than the proposed model by Bevendorff *et al.* [2019], which proves the power of deep neural networks on authorship-based tasks.

Sari *et al.* [2017] used continuous representations through a neural network along with the classification layer to figure out who wrote something, while citetshrestha2017convolutional used CNNs to figure out who wrote short-digit text (like tweets) by estimating the importance of input text fragments to make the model easier to understand. Zhang *et al.* [2018b] applied a novel strategy to encode the syntax parse tree of the sentence into a learnable distributed representation. Specifically, they build an embedding vector for each word in the text by encoding the path as a syntax tree corresponding to the word. The work by Sari *et al.* [2018a] is one of the state-of-the-art feature-based methods that extracts features using various stylometric features and achieves excellent performance. However, it does not take full advantage of the semantic features. However, it does not take full advantage of the semantic features. Yao *et al.* [2019] is the convolutional graph network for text classification but cannot run on IMDb62 datasets because a huge text graph takes up too much computer memory.

Jafariakinabad *et al.* [2019]; Jafariakinabad and Hua [2019] looks at how strong and reliable the syntactic recurrent neural network is at storing a document's syntactic patterns in a hierarchical way for AA tasks. However, it is unclear what the learned vectors are capturing. The BLSTM-CNN with subword embedding based on the BPE algorithm [Sennrich *et al.* 2016] helps us understand and record syntactic patterns better. When compared to the syntactic recurrent neural network, the BLSTM-CNN with subword embedding based on the BPE algorithm can understand and record more syntactic patterns. Additionally, it overcomes the limitation of memory usage that prevents the convolutional graph network from running on IMDb62 datasets.

5.3 PROPOSED METHOD

The first portion of our system, as presented in Figure 36 adopts a byte-pair-encoding (BPE) algorithm in the text representation as an embedding tactic to transform the pure text into numerical representations. In the feature extraction as the second phase, we feed the embedding modules into a bidirectional LSTM to understand the underlying semantics and apply CNN max-pooling over time to capture the local spatial syntactical position on writing style from the input text. The classification phases consist of a fully-connected layer and soft-max function, which is sufficient to fit the function which takes the features and outputs the classification result. Besides, we combined annealed Gaussian noise with training the model to learn the writing style representations for AA tasks, which helped avoid overfitting and achieved lower training loss. We evaluated the model's performance using k-fold cross-validation with a large spectrum of public datasets for authorship attribution on CCAT50, IMDb62, Blog50, and the newly assembled Twitter dataset for the AA tasks.

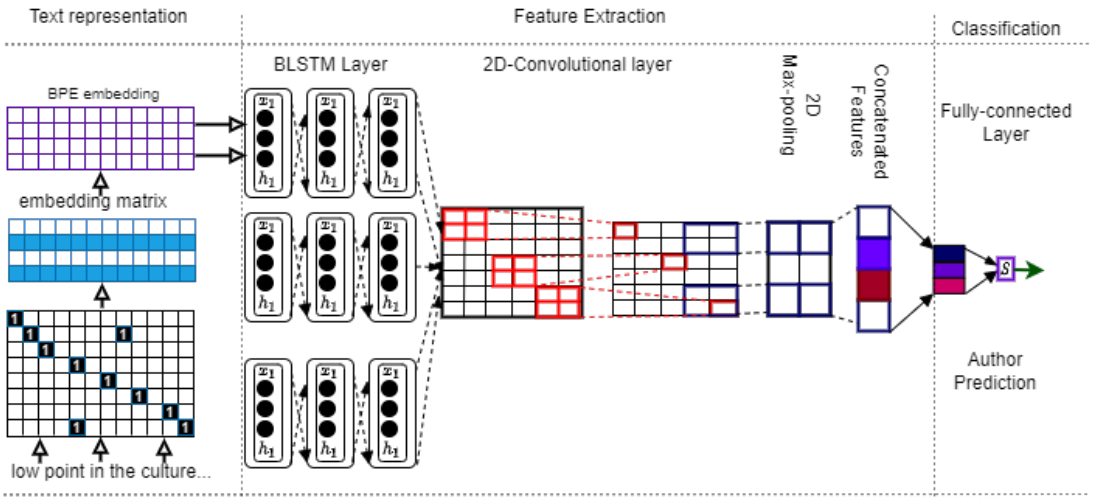


Figure 36: BLSTM-CNN model with subword information

5.3.1 *Text Representation*

The first part of the proposed system is text representation based on character-level encoding that employs byte-pair encoding (BPE) to convert subwords in the pure text of the authors into numeric representations. BPE brings the perfect balance between character- and word-level hybrid representations, which makes it capable of applying the proposed model to large unstructured corpora like social media text. BPE behavior also enables the encoding of any rare words in the corpus with appropriate subword tokens without introducing any "unknown" tokens into the model.

5.3.1.1 *Sub-word embedding*

BPE is a tokenization technique adopted in machine translation to deal with imaginary word problems or hidden writing in a given text. It is unsupervised and requires no information about the author. The algorithm of BPE first initializes a symbol by splitting the input text into characters. Then, iteratively count all symbol pairs and replace each occurrence of the most frequent pair *(x,y)* with a new *xy* symbol and add it to the symbol set named "merge operation". Each merge operation generates a new symbol. The size of the final symbol set is equal to that of the first single character, plus the number of merge-operations. The only hyperparameter of BPE algorithm, as shown in Algorithmus 3 is the number of merge operation and if the number of merge operation is large, we will get a better vocabulary.

Algorithmus 2 : Computation of replacing string

Input : training data D of words split into character sequence with number N of rules
Output : list of K of N merge rules

$K := [\,]$
while length $(K) \leqslant$ N **do**
  $(x, y) := \underset{(x,y)}{\operatorname{argmax}}\, count_D(x, y)$
  $rule := \langle (x, y) \longrightarrow (xy) \rangle$
  $D := apply(rule, D)$
  $K := append(rule, K)$
return K

For untokenized unique text, we first split it into a single character and then iteratively do the merge operations following the merge order in the training step; until there are no more symbols that can be merged. That is to say, if the number of merge operations is large, the token will tend to have more characters, and the granularity tends to be large. Otherwise, the granularity of the original text will be small. In our system, we do not use BPE as a compression algorithm. Instead, we use this algorithm to find sub-words as n-grams with high frequencies for word segmentation, achieved if we joined characters together. However, we did not substitute them with new symbols. An example of how subwords are obtained from a raw input text after $\mathcal{N}$ iterations is shown in Table 9.

The text has now been subdivided into subword sequences. To use subword embedding to represent the text, we first create a one-hot vector for each subword type. The one-hot vector for the ith subword in the vocabulary is a sparse binary vector o_i which has 1 as the ith element and all 0 for others. After that, we project this embedding hyperspace onto a smaller hyperspace by multiplying the one-hot embedding with a subword embedding matrix S with size $N \times D$, where N represent the sub-word vocabulary size and D is the dimension for the target embedding hyperspace. Therefore, we represent each sub-word information as a dense vector $s_i = S_o^T i$, and the text with length T is represented by a sequence of subword embedding vectors $(s_1, s_2, \cdots, s_T)$. Therefore, the subword embedding matrix S are trained together.

5.3.2 *Feature Extraction*

The second part of our system is feature extraction based on the sub-word embeddings feeds into BLSTM and CNN module, and the final output was a neuron representing the probability of feature vectors belonging to certain authorship. Given the subword embedding, we use BLSTM to find the inherent grammatical relationship in the author's writing style embedded in the subword and the CNN to capture the local syntactical information from the input text.

Table 9: BPE example given raw input text.

Iteration	Words
0	workers work in workshop
1	workersworkinworkshop
2	*wo*rkers*wo*rkin*wo*rkinshop
3	*wor* kers*wor* kin*wor* kshop
4	*work* ers *work* in *work* shop
N	*work er* s *work in work shop*

5.3.2.1 *Bidirectional LSTM Layer*

Long short-term memory (LSTM) was first proposed by [Schuster and Paliwal 1997] to overcome the gradient vanishing problem of recurrent neural network (RNN). The idea is to introduce an adaptive gating mechanism, which decides the degree to keep the previous state and memorize the extracted features of the current input text. So, given a sequence of input text $X = \{x_1, x_2, \cdots, x_l\}$, where l represent the length of the input text, LSTM processes the input as a subword information. At each time-step t, the memory c_t and the hidden state h_t are updated as following:

$$\begin{aligned} f_t &= \sigma(x_t W_{x,f} + h_{t-1} W_{h,f} + b_f) \\ i_t &= \sigma(x_t W_{x,i} + h_{t-1} W_{h,i} + b_i) \\ o_t &= \sigma(x_t W_{x,o} + h_{t-1} W_{h,o} + b_o) \end{aligned} \tag{15}$$

where $W_{x,f}$ and $W_{h,f}$ are the weights of the LSTM from input to forget gate and from the hidden state to the forget gate. The $W_{x,i}$ and $W_{h,i}$ are the weights from input to input gate and from the hidden state to the input gate. Similarly, $W_{x,o}$ and $W_{h,o}$ represent the weights from input to output gate and from the hidden state to the output gate. Finally, b_f, b_i, and, b_o are the bias for the forget, input and output gate. σ denotes the logistic sigmoid function. The memory cell can be calculated as:

$$\hat{c}_t = \tanh(x_t W_{x,c} + h_{t-1} W_{h,c} + b_c) \tag{16}$$

where $W_{x,c}$ and $W_{h,c}$ represent the weights of the LSTM from input to the memory and from the hidden state to the memory, respectively, and the b_c denotes the bias. The memory cell at the time step t was computed as follows:

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \tag{17}$$

where $\otimes$ denoted element-wise multiplication. The hidden state can be updated as:

$$h_t = i_t \odot \tanh(c_t) \tag{18}$$

The RNN model was forward and the output at the time steps t depends on the past context as well as the hidden state, e.g. the future context. For example, Schuster and Paliwal [1997] introduced a BLSTM to extend the unidirectional LSTM by introducing a second hidden layer, where the hidden connections flow in opposition temporal order. Therefore, the model is able to exploit information from both the past and the future. In this study, BLSTM is used to capture the past and the future writing style information. As shown in Figure 35, the system contains two sub-networks for the forward *(f)* and backward *(b)* sequence context based on the subword embedding from the input text at each time step t as follow:

$$\begin{aligned} h_t^f &= \sigma(x_t W_{x,h}^f + h_{t-1}^f W_{h,h}^f + b_h^f) \\ h_t^b &= \sigma(x_t W_{x,h}^b + h_{t-1}^b W_{h,h}^b + b_h^b) \end{aligned} \tag{19}$$

The output at each time t can be computed as:

$$h_t = [h_t^f \otimes h_t^b] W_{h,o} + b_o \tag{20}$$

5.3.2.2 *Convolutional Layer*

From the BLSTM layer, we have access to the future context as well as the past context, o_t is related to all other writing style (or words) in the given text. In this study, we effectively treat the matrix as feature vectors, so 1D convolution and the max pooling operation were used to capture local syntactical information.

For matrix $H_t = \{h_1, h_2, \cdots, h_l\}$, $H \in \mathbb{R}^{l \times d^w}$ obtained from BLSTM layer, where d^w is the size of the subword embeddings. Then narrow convolution is utilized to extract local features information over H. The convolution operation involves a filter $m \in \mathbb{R}^{k \times d}$, which is applied to a window of k subwords and d feature vectors. For instance, a feature $o_{i,j}$ is generated from a window of vectors $H_{i:i+k-1,j:j+d-1}$ as

$$o_{i,j} = f(\boldsymbol{m} \cdot H_{i:i+k-1,\, j:j+d-1} + b) \tag{21}$$

where i ranges from 1 to (l-k+1), j ranges from 1 to $(d^w - d + 1)$, $(\cdot)$ represents dot product, $b \in \mathbb{R}$ is a bias and an f is a non-linear function similar to hyperbolic tangent. So, we applied the filter to each possible window of the BLSTM layer matrix H to obtain a feature map $\mathbf{O}$:

$$O = [o_{1,1}, o_{1,2}, \cdots, o_{l-k+1,d^{w-d+1}}] \tag{22}$$

with $O \in \mathbb{R}^{(l-k+1)\times(d^{w-d+1})}$ represent one convolution filter process. The convolution layer have multiple filters for the same size filter to learn complementary features, or multiple kinds of filter with different size. Then 2D max pooling operation $p \in \mathbb{R}^{(p_1 \times p_2)}$ is utilizes to obtain a fixed length vector by applying it to each possible window of matrix O to extract the maximum value:

$$p_{i,j} = down(O_{i:i+p_1, j:j+p_2}) \tag{23}$$

where $down(\cdot)$ represents the map pooling operation function, $i = (1, 1+p_1, \cdots, 1+(l-k+1/p_1-1)\cdot p_1)$ and $j = (1, 1+p_2, \cdots, 1+(d_w - d + 1/p_2 - 1)\cdot p_2)$. Then, we compute the pooling operation as follows:

$$\begin{aligned} h^* = [&p_{1,1}, p_{1,1+p_2}, \cdots, p_1 + (l-k+1/p_1-1)\cdot p_1, \\ &1 + (d_w - d + 1/p_2 - 1)\cdot p_2] \end{aligned} \tag{24}$$

where $h^* \in \mathbb{R}$, and the length of h^* is $\lfloor l-k+1/p_1 \rfloor \times \lfloor d_w - d + 1/p_2 - 1 \rfloor$.

5.3.2.3 *Classification Layer*

For the AA tasks, the output h^* from the max pooling was passed over the fully-connected layer of the input text X, then feeds it to a softmax function as a classifier to predict the inherent writing style related to a particular candidate author $\hat{y}$ from the set of discrete set of author (or classes) Y. So, the classifier takes the hidden state h^* as input as follows:

$$\hat{p}(y|x) = softmax\left(W^{(x)}h^* + b^{(x)}\right) \tag{25}$$

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, \hat{p}(y|x) \tag{26}$$

To learn the model parameters, we minimize the cross-entropy loss as the training objective by calculating the loss as a regularized sum:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(y_i) + \lambda\|\theta\|_F^2 \tag{27}$$

where $\boldsymbol{t} \in \mathbb{R}^m$ represents the one-hot encoding for the ground truth values, $\boldsymbol{y} \in \mathbb{R}^m$ is the predicted probability of the candidate author by softmax, m is the number of expected target authors, and λ is the ℓ_2 regularization parameter and we train with Adam optimization algorithm [Kingma and Ba 2014]. Finally, the pseudocode of the proposed model is given in Algorithmus 3, where we use simplified variables to make the procedure clear.

Algorithmus 3 : Pseudocode for BLSTM-2DCNN max-pooling with subword information

Input : Training data $\mathbf{X} = \{X_i\}_{i=1}^{n}$ m represent the batch size, n is the number of training samples, w is model parameters and l represent the length of the input text.

Output : trained model.

3.1 Randomly initialize w;

3.2 **foreach** *each iteration* **do**

3.3 **forall** $k \in \{1, 2, \cdots, \lfloor \frac{n}{m} \rfloor\}$ **do**

3.4 Sample from each batch $\mathbf{X}_k$ from $\mathbf{X}$;

3.5 Divide each sample in the batch into the sequence $\{\mathbf{X}_k^1, \mathbf{X}_k^2, \cdots, \mathbf{X}_k^l\}$;

3.6 Feed sequential batch into BLSTM consisting of forward and backward neuron, respectively, and outcome two output sequence $\{\boldsymbol{h}_{fk}^1, \boldsymbol{h}_{fk}^2, \cdots, \boldsymbol{h}_{fk}^l, \boldsymbol{h}_{bk}^1, \boldsymbol{h}_{bk}^2, \cdots, \boldsymbol{h}_{bk}^l\}$ Equation 19;

3.7 Concatenate the BLSTM layer to obtained $H \in \mathbb{R}^{l \times d^w}$, the narrow convolution is used to extract local dependent features over H to produce a feature map O ;

3.8 Then, $p \in \mathbb{R}^{(p_1 \times p_2)}$ is applied to each possible window of matrix O to obtain h^* Equation 24 represent the stylometric representation (e. g. writing styles) of the input $\mathbf{X}$;

3.9 Feed the output h^* into the softmax classifier layer and obtain the classification result in Equation 25 and Equation 26 respectively;

3.10 Update w by minimizing with the cross-entropy loss in Equation 27 using Adman algorithm [Kingma and Ba 2014];

5.4 EXPERIMENT AND RESULTS

We address the datasets, hyperparameters, and promising outcomes of the proposed BLSTM-2DCNN Gaussian noise with subword embedding.

5.4.1 *Datasets*

We evaluated the proposed model on CCAT50, IMDb62, Blog50, and new Twitter datasets as discussed in Chapter 3.

5.4.2 *Hyperparameters tuning and training*

For all the datasets, we used Adam optimization [Kingma and Ba 2014] with a mini-batch size of 64, an initial learning of 0.001, a ReduceLRonPlateau schedule with 5 epochs and a decay factor of 0.5. During training, the dimension of the subword vector is 300. The hidden units of BLSTM are 128. We use 100 convolutional filters for the window sizes of (3,3) with a max-pooling size of (2,2). For regularization, we employ Dropout operation [Hinton *et al.* 2012b] with a dropout rate of 0.5 for the subword embeddings, 0.3 for the BLSTM and 0.2 for the penultimate layer with Gaussian Noise of 0.2 active at the training time. We also use ℓ_2 penalty with coefficient 10^{-5} over the parameter and trained for 20 epochs. All word vectors and feature vectors are randomly initialized and learned, and updated during the training process. The dimensions of the word vector and hidden layer size are d=64 in all models. We use 128 convolutional filters, each for window sizes of (3,3) and 2D pooling sizes of (2,2). Our models were trained on the graphics processor (GPU), which leverages NVIDIA GeForce 2080Ti. All the experiments were repeated 5-fold cross-validation, and the accuracy reported represents the mean average of the prediction writing style of the authors.

5.4.3 *Results*

The overall authorship attribution accuracies of our methods and the baseline are provided in Table 10. The missing values denoted as "(–)" in Table 10 indicate that the feature and the model were not considered by the previous SOTA methods. The proposed architecture in Figure 35 contains three different class of deep learning such as BLSTM-2DCNN, BLSTM-2DCNN word embedding and BLSTM-2DCNN gradient noise with subword information. Thus, Table 10 presents the performance of all three models and other state-of-the-art models on four datasets for authorship-based tasks.

Table 10: Performance comparison and accuracy scores on four mainstream AA datasets.

Models	**CCAT50**	**IMDb62**	**Blog50**	**Twitter50**
Imposters [Koppel *et al.* 2011]	–	76.90	26.00	52.50
SVM frequent [Plakias and Stamatatos 2008]	67.00	81.40	–	–
LDAH-S with topics [Seroussi *et al.* 2011]	–	72.00	18.30	38.30
SVM affix [Sapkota *et al.* 2015]	69.30	69.90	–	–
Style, content & hybrid [Sari *et al.* 2018a]	–	75.76	84.51	–
CNN n-gram [Shrestha *et al.* 2017]	76.50	95.21	53.09	–
Cont. n-gram[1] [Sari *et al.* 2017]	72.60	95.12	52.82	–
Syntax-CNN [Zhang *et al.* 2018b]	81.00	**96.16**	56.73	–
Cont. CNg + NF[2] [Verma and Srinivasan 2019a]	80.07	97.89	–	–
Document embedding [Parikh *et al.*]	72.88	95.93	–	–
SRNN [Jafariakinabad and Hua 2019]	82.35	–	**61.19**	–
BertAA [Fabien *et al.* 2020b]	–	90.70	59.70	–
BLSTM-2DCNN	73.25	81.25	48.15	49.52
BLSTM-2DCNN word embedding	74.50	89.72	53.26	79.30
BLSTM-2DCNN subword+Gaussian noise	**83.42**	93.72	60.67	**87.76**

As shown in Table 10, the BLSTM-2DCNN+Gaussian noise with subword embedding achieves comparative performance on three out of four datasets. Gaussian noise was combined with ℓ_2 regularization to gain roughly 10% better performance when compared to both traditional methods and the existing CNN based models. Essentially, it achieves 2.9%, 1.6% and 0.9% test accuracy on CCAT50, Blog50 and Twitter datasets, respectively. In addition, the performance of the proposed model is superior to that of the CNN and BertAA model [Ruder *et al.* 2016a; Fabien *et al.* 2020b], which shows that learning from characters or leveraging on the pre-trained language model without feature engineering task can help to improve the performance for AA tasks. Our method is much better than the BertAA model, which validates the effectiveness of integrating a pre-trained BERT [Devlin *et al.* 2018a] language model with an extra dense layer to perform authorship classification. In addition, different from existing CNN-based methods, we leveraged the extracted features employing the BPE algorithm to represent words (e.g., the writing style) by its index in the vocabulary together with its subword vector classes. Consequently, the proposed model inherits the advantage of both traditional CNN-LSTM model [Ruder *et al.* 2016a; Gupta *et al.* 2017; Jafariakinabad *et al.* 2019] and BPE algorithms with Gaussian noise, which contributes to the performance improvement for the AA dataset as shown in Table 10.

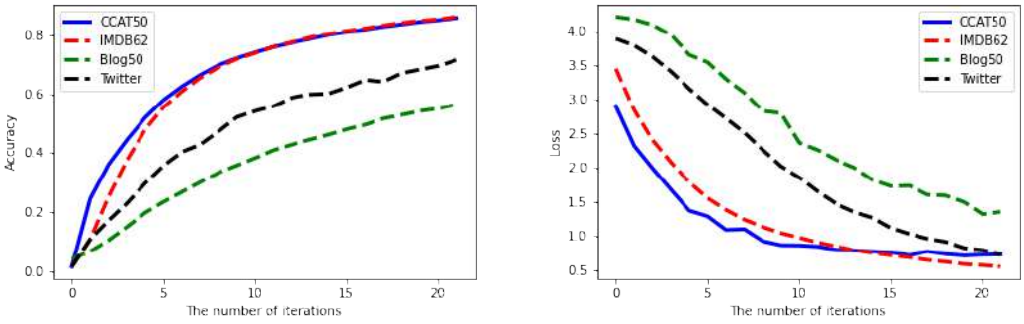
The proposed BLSTM-CNN model with subword information outperforms all the CNN variants by more than 10% and 4% for both Blog and Twitter datasets with 50 authors. Differences for the IMDb62 domain render less discriminatory words or character sequences

Table 11: Performance comparison of CNN and BretAA model on all the datasets.

Models	CCAT50	IMDb62	Blog50	Twitter50
CNN-char [Shrestha *et al.* 2017]	76.50	95.21	53.09	84.29
Document embedding [Parikh *et al.*]	77.03	92.30	62.40	82.50
[Verma and Srinivasan 2019a]	80.07	**97.89**	61.27	73.40
BertAA [Fabien *et al.* 2020b]	78.90	93.00	64.40	62.50
Our method	**83.42**	93.72	60.67	**87.76**

when authors review similar movies. The BertAA model is boosted on IMDb62 because they are less sensitive to topical divergence. They are, however, less helpful in short digit text, e. g. Blog50 and Twitter domain, where hashtags or emoticons are the most characteristic features.

To further substantiate the effectiveness of our model, we tested CNN and BertAA models and the Gaussian noise, respectively. We then reported the performance of the results in Table 11. For the CNN-based and BertAA model, we add Gaussian noise before the softmax classifier on the same network structure. Comparing CNN-based and BertAA models, we see that each model can improve authorship classification accuracy using the same extracted features from BPE algorithms. In addition, it can be seen in Table 11 that our model is superior to the counterpart CNN-based or BertAA model with a pre-trained weighted vector. The accuracy and convergence curves (e. g. loss) on the datasets were displayed in Figure 37, respectively. However, in Figure 38, we can see the best classification accuracy with faster convergence speed in the training process for Twitter datasets compared to the CNN-char model [Ruder *et al.* 2016a].

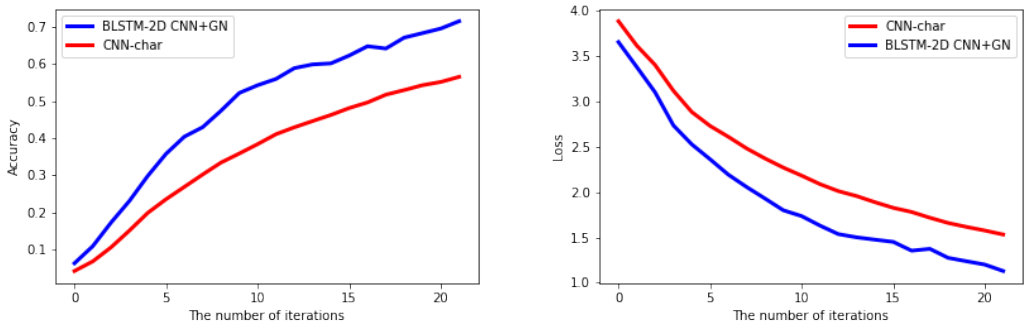


(a) Accuracy (b) Loss

Figure 37: The accuracy and convergence curve over iterations on all the datasets

5.5 DISCUSSION

The best performance result for each SOTA method is reported because the source codes of the comparative methods are unavailable. The proposed method is seen to significantly outperform most of the investigated methods, including the deep learning methods by an accuracy margin of 1.07%, and 0.96% on the CCAT50, and Twitter datasets, but less superior on the IMDB62 and Blog50 dataset by 2.44% and 0.32%. The accuracy on the CCAT50 dataset with the most frequent 3-gram [Plakias and Stamatatos 2008] did not perform well with the affix and punctuation 3-grams features reported by [Sapkota *et al.* 2015]. In contrast,



(a) Accuracy (b) Loss

Figure 38: The accuracy and convergence curve over iterations on Twitters dataset.

Style-HAN method that encode syntactic patterns of the text using only their corresponding part-of-speech tags perform comparatively on the CCAT50 dataset but have negligible results with Blogs datasets, demonstrating that the method based on sub-word information are competitive to capture relevant text representations for the AA task. CNN-word, CNN-word-word achieve comparatively worse performance for short-message domains such as Blog and Twitter. They, however, provide a better performance in domains in which topical information is not discriminatory, such as movie reviews. We suspect that CNN-character and word stores discrete n-grams for each author, making it very simple to distinguish a topic preference for each author but not satisfactory to capture the writing style of a different author or tweet written by the same author. If such a correlation exists, the lower accuracy reached by the CNN model [Ruder *et al.* 2016a; Jafariakinabad and Hua 2019], and Syntax-CNN [Zhang *et al.* 2018b] prove that the model carries less topical information, and the stylistic representation loses cohesiveness when the size of the author is too large. Still, our proposed model is superior in capturing better stylistic information such as words, topics, punctuation, or phrases re-posted by some authors when publishing their comments.

Moreover, the shallow neural network (that is FastText embedding) [Sari *et al.* 2017] beat the SVM [Plakias and Stamatatos 2008; Sapkota *et al.* 2015] with similar feature sets, while our method shows superior encoding ability for AA tasks on all the datasets. The fine-tuned BertAA [Fabien *et al.* 2020b], and continuous n-gram representation Sari *et al.* [2017] with pre-embedding vectors can better identify author writing style. These method are sensitive to topical writing divergence between authors, but our method get boosts with 2D filter and max pooling size to get better performance for AA tasks. Besides, our methods outperform SVM with frequent 3-grams, LDAH-S, Imposters, CNN variants, and other deep neural network methods on 2 out of 4 datasets in Table 10 in accuracy except for the CCAT50 and newly Twitter dataset. The result demonstrated the aptitude of our methods to handle the class variation problem inherent in the real-world application [Stamatatos 2009b].

5.6 CHAPTER SUMMARY

Chapter 5 demonstrates that input-embedded vectors employing subword in-formation feed with the BLSTM-2DCNN model could learn stylometric representations of different linguistic modalities for AA tasks. It showcases such a configuration's effectiveness in dealing with common spelling errors from unstructured texts due to orthography and phonetic reasons, then learns stylistic and topical information to classify the author. In addition, the Gaussian noise is introduced to the fully conventional layers, which substantially reduces the large number of parameters arising from the model structure. Thus, the convergence rate of the model significantly speeds up and improves the classification accuracy. We evaluated the model against the state-of-the-art methods for an extensive range of authors, demonstrating the proposed model's effectiveness in handling morphological variance and its applicability across authorship-related tasks. In the future, it is hoped to explore combining the model with

a self-attention mechanism to model different linguistic levels (e.g., structure, POS tagging, dependency, and semantics) applying subword information to improve the alignment of words in the input texts during training to advance the research on authorship-based tasks. We would like to conduct further analysis of different variations of authors to improve the performance of the proposed method.

Part VI

POST-AUTHORSHIP ATTRIBUTION USING ATTENTION-BASED BIDIRECTIONAL GATED RECURRENT UNITS WITH CONVOLUTION

6

POST-AUTHORSHIP ATTRIBUTION USING ATTENTION-BASED BIDIRECTIONAL GATED RECURRENT UNITS WITH CONVOLUTION

Stylometric is a quantifiable study of written style that uses computer-generated text, online text samples, or extensive text samples. It is based on the assumption that every individual has a distinct, consistent, and recognised writing voice. As a result, by examining these stylometric features, it is feasible to identify the author of a given text. By examining these stylometric features, it is feasible to identify the author of a given text with a certain degree of accuracy. It has been applied in a variety of sectors, including the detection of misogynistic, xenophobic, and abusive writings for cybercrime, as well as literary science, marketing politics, and socialisation. It has also been used for detecting plagiarism and forensic linguistic analysis. The major problem is the potential bias and inaccuracies of the manually extracted features, which have been time-consuming and labour-intensive and have very limited generalizability, not to mention the difficulty of interpreting the features that contribute to the actual identification of the author of the suspected text. Typically, the text content of the online text snippet is short with an obfuscating vocabulary, and the terminology used in the content includes idioms, onomatopoeias, homophones, phonemes, synonyms, acronyms, anaphorics, and polysemy. The challenge is how to represent each unit of textual data as a numeric vector and determine an author's writing style that can adequately characterise a writer's unique voice and distinguish it from others. In this chapter, we explore the amalgamation of pre-trained word embedding vectors to extract syntactic and semantic features to learn linguistic-motivated elements in the text. Motivated by a fully convolutional neural network and sequence-to-sequence model in the field of natural language processing algorithms, which have been used to substantially reduce the number of parameters in the text classification. To get deep features, a pre-trained embedding model is used. These are then fed into bi-directional long-term (BiLSTM) and gated recurrent units (BiGRU) to help learn semantic features in text in a way that makes sense. The number of dimensions of the features is cut down after the feature vector is fed into an attention-based, convolutional, and pooling mechanism. The writing style is then learned without regard to position. A softmax function is used for feature classification, and the proposed model is evaluated against 20 trend-setting methods using four benchmark experimental datasets. Results demonstrated its effectiveness compared to the methods on two datasets while producing comparable effects on the remaining two. The ABC method showed promising results in feature classification, suggesting its potential for further development and application.

6.1 INTRODUCTION

The increasing number of computerised information systems, personal devices, and globalisation of the Internet and social networking sites has fundamentally transformed the manner in which humans generate and digest information for profitable ventures [Ding *et al.* 2017; Modupe *et al.* 2022]. Countless pieces of online text snippets and documents, such as news, blogs, and tweets, are generated on a daily basis as a source of entertainment [Whiting and Williams 2013], advertisements [Yun *et al.* 2020], education [Burbules 2016; Manca 2020] and many other valuable venture [Modupe *et al.* 2022]. In every millisecond, new information in the form of stories from news websites, posting links to a news story, commenting on new ideas posted by a scientific research group e. g.collaboration [Greenhow and Li 2012], sentiments, and emotions via social media [Ding *et al.* 2017; Modupe *et al.* 2022]. These forms of content creation and sharing have become an integral part of our daily lives [Mihailidis 2014] shaping our opinions, influencing our decisions, and connecting us with

others around the world [Naeem 2021]. With the rapid advancement of technology, the volume and variety of information being produced and consumed continue to grow exponentially.

The proliferation of online text snippets and documents from social media like Twitter posts, Facebook status updates, short message service (SMS) messages, or chat conversations presents both opportunities and challenges. On one hand, it allows for greater access to diverse perspectives and knowledge. On the other hand, these posts are improperly utilised to promote illegal activities, such as pirated software, terrorist communications, child pornography, harassment, abusive, obscene, menacing, corrosive, and antediluvian messages [Modupe *et al.* 2011b]. This raises serious concerns about the negative impact of social media on society. For example, tweets were posted to several individuals to influence the 2016 US presidential election, and almost 65% of the tweets were found to be blatantly lying or included unsubstantiated opinions, which was coded as fraudulent news and had a significant impact on public opinion and voting behaviour [Padda 2020]. This manipulation of information through social media platforms highlights the potential power and influence that fraudulent stories can have on society and political outcomes. In this era where misinformation spreads rapidly, it underscores the need for revealing the identity and sociolinguistic characteristics of an underlying writer. Identifying the sources of false information is crucial in order to combat the spread of misinformation and protect the integrity of public discourse. Additionally, people hide their identities when posting sensitive information [Wängqvist and Frisén 2016]. The juvenile offenders usually mask their identities across multiple networks to circumvent the prowess of text mining methods [Altakrori *et al.* 2018]. Using proxy Internet addresses and hashtags, mentioning people in texts, and reposting texts written by others is a common phenomenon [Alonso-Fernandez *et al.* 2021]. The discussion of a true sender can be forged through an anonymous server to dispense offensive text messages across diverse pseudonymous channels [Rocha *et al.* 2016]. The anonymity of social media posts also imposes unique challenges and risks of tracing the identity of an offender without compromising privacy in cyberspaces [Modupe *et al.* 2022].

As post-authorship attribution and forensic methods based on stylometry become more popular, it is now possible to use the author's language as a fingerprint to quantify the writing style of an author [Ding *et al.* 2017; Modupe *et al.* 2022]. The identification of the perpetrator's, which may be associated with their true identity, has become critical for investigations and legal proceedings [Venčkauskas *et al.* 2015; Rocha *et al.* 2016; Alonso-Fernandez *et al.* 2021]. None of the relevant works on short texts via social media focus on the ability to interpret and visualise [Zheng and Jin 2023; Wu *et al.* 2021a] the contributing factors that determine an offender's writing style [Rocha *et al.* 2016; Shao *et al.* 2019; Belvisi *et al.* 2020; Alonso-Fernandez *et al.* 2021]. However, the major challenge for post-authorship attribution (PAA) is finding a suitable writing style adequately characterising an author [Modupe *et al.* 2022].

To make PAA systems more accurate, we wanted to create a bidirectional attention-based long-term short memory and a bidirectional gated recurrent unit. These units would also have convolution and pooling mechanisms built in. The unique contributions of this study to theory and practice are the following:

- The amalgamation of a pre-trained model for extracting stylometric features improves the authorship attribution system.
- Introducing an attention-based bidirectional long short-term memory (BILSTM) and bidirectional gated recurrent unit integrated with convolution and pooling mechanisms used to train amalgamation of a pre-trained language model.
- Introducing an interactive system to visualise the results of the PAA method could be useful in an evidence-based forensic investigation.
- Demonstrating the performance of the proposed ABC method through experimental comparison with existing prominent methods.

Chapter 6 is succinctly organised as follows: Section 6.1 provides the introductory message. Section 7.2 discusses the related studies in chronological order. Section 6.3 presents the theoretical background of the proposed ABC model, and Section 6.4 designates the proposed method. Experimental results, including datasets, parameter settings, discussion of results, and visualised results, are explained in Section 6.5. In Section 6.5.5, we provide a conclusion and future work.

6.2 RELATED WORK

Research on the authorship attribution process has been exhilarating for a long time. Several studies present various methods of developing automated systems for identifying the true author of an anonymous online text snippet. Recently, researchers focused on online text snippets attributable to the growth of web applications and social networks [Alonso-Fernandez *et al.* 2021]. In most cases, solutions in earlier authorship attribution studies typically comprise two stages. The first stage relates stylometric features, including word level (lexical), punctuation and function words (syntactic), sentence and paragraph (structural), topic (content), and the unique elements of a particular author (idiosyncratic). These are extracted from each unit of textual data as a numeric vector to characterize the unique writing style of a particular author [Stamatatos 2009b].

Lexical features are statistical measures of text characters and word-based lexical text variations, such as vocabulary richness and word length distribution [Stamatatos 2009b]. Syntactic features capture the universal part of speech (POS) tags from the text structure [Alonso-Fernandez *et al.* 2021]. Structural features are associated with the organization of texts, such as the average number of words in a sentence, lines, punctuation, and indentations or paragraphs [Hirst and Feiguina 2007]. Content features explore, for example, the distribution or frequency of specific keywords. Last, the idiosyncratic features capture the particularities of an author concerning misspelled linguistic words, abbreviations employed, or additional special characters (such as emojis) [Alonso-Fernandez *et al.* 2021].

The second stage of the authorship attribution process is allied with handling the texts per author, employing two methods. First, the profile-based method concatenates all identified texts from an author into a single text file where the author's writing style is deduced. This is conducted by using probabilistic methods, such as naïve Bayes or distance measure [Dunn *et al.* 2016]. Second, instance-based methods extract features from each text separately and train them with machine learning classifiers, such as support vector machines (SVM), decision trees (DT), or random forest (RF) to identify the author [Belvisi *et al.* 2020; Potha and Stamatatos 2014]. This method requires multiple training samples per author to develop an accurate model, which may not always be available in forensic examination [Venčkauskas *et al.* 2015].

The connotation of feature extraction is vital to authorship attribution. Past studies [Neal *et al.* 2017b; Rocha *et al.* 2016] applied character and word n-grams coupled with the number of words, short phrases, sentence length, digit frequency, and others from online text snippets to identify authors. To handle offensive posts on the Internet or social media, [Okuno *et al.* 2014a] reported an accuracy of 53.2% in identifying the authorship of roughly 10,000 Twitter users (90 for training and 30 for testing). They used POS-tag combined with n-grams to control the weight of N-gram features, and the execution time, and to reduce the number of features. Belvisi *et al.* [2020] applied stylometric features, characters, and word n-grams, with a self-acquired database of 40 users and 120-200 tweets per user.

The number of tweets per user was in the same range; however, the number of users was dissimilar, which explains why the idiosyncratic feature achieved a higher accuracy of 98.5% for 40 users [Marcu 2000]. The idiosyncratic features cannot learn and generate an author's writing style in tweets but only capture individual eccentricities. Neal *et al.* [2018]used short

blocks of 50-100 characters from blog posts to conduct continuous user verification. They used stylometric features and character n-grams with a total dimensionality of 671. They fed it with isolation forest (IF) to train 3–10 blocks of text and assess them with the subsequent block (to simulate continuous sessions). The purpose is to mimic login sessions where the user is verified as they produce a new text. In contrast to regular verification, machine learning is trained to classify test samples as normal or abnormal and achieve an accuracy exceeding 98%.

The current research method is helpful for natural language processing (NLP), where learning and representing an author's writing style to discuss the anonymity of a questioned text used for crime during a pretrial case is challenging. With the growing number of criminals using the Internet or social media to distribute a wide range of illegal materials and false information globally anonymously. Yang and Chow [2014a] used parts-of-speech and function words to extract content-based and style-based features to select 2000 features with the most significant information gain (IG). They obtained 20 potential authors from 2000 candidate authors using logistic regression (LR) to achieve a high-precision rate of 80.15% on age and 85.1% on gender attribution. Koppel *et al.* [2011] used character n-grams to capture the stylistic and morphological information of the authors. They employed the similarity-based method to attribute 500-word snippets to 1000 authors to achieve a high-precision rate of 93.2%. Seroussi *et al.* [2011] demonstrate the importance of the nearest-neighbour collaborative approach and traditional topic modelling, such as Latent Dirichlet Allocation (LDA) [Blei *et al.* 2003], using Hellinger distance between the LDA topic distributions for an unknown text and an author's profile as a measure for authorship on the IMDb62 datasets [Seroussi *et al.* 2010].

The survey of a variety of feature types and categorization methods proposed in the past for authorship attribution based on literature corpus, email, and blog corpora was reported [Koppel *et al.* 2009]. The SVM and Bayesian regression methods were experimented with to achieve 80% on the 1000 most frequent words and 86% on 1000-character n-grams with the highest IG. Peng *et al.* [2003a]; Stamatatos [2013] identified word and character-level n-grams as the most compelling features for identifying authors. Schwartz *et al.* [2013b] used the K–signature with character and word n-grams to interpret the unique writing style, appearing in the least K% tweets posted by an author. They achieved an accuracy of 50.7% on a dataset holding 50 authors with 50 training tweets per author. An accuracy of 71.2% was achieved with 10,000 training tweets per author. The advantage of using character-level N-gram features was attributed to the high priority of employing affix and punctuation as a subgroup of word features to capture morphological information for authorship attribution [Sapkota *et al.* 2015].

Rocha *et al.* [2016] presented an exhaustive bibliography survey of the state-of-the-art (SOTA) techniques for PAA tasks. They employed shallow classifiers, such as SVM and RF allied with a bag-of-words (BOW) feature representation. They explored the performance dependency of their proposed method with multiple authors and messages; however, their results are still far from satisfactory, despite advancing the SOTA for PAA tasks. The authors presented a classification accuracy of less than 65% in a new experimental dataset of 50 authors. This number plummets to less than 45% in 1000 authors, indicating that PAA is still an open research endeavor that requires further improvement. Most of these methods rely on discrete feature representation, which suffers from data sparsity, disregarding the semantic relationship between the vector representation for authorship attribution.

In recent years, advancements in neural networks (NN) allied with the emergence of more affordable and better-powerful graphics processing units (GPU) and the tremendous data availability culminated with the rise of what is now called deep neural network (DNN) [Goodfellow *et al.* 2016; LeCun *et al.* 2015]. The term refers to several connections of network layers and a vast number of parameters to develop a robust model capable of solving challenging problems relying upon linear models over sparse feature vectors with high dimensionality systems for an extended period. This novel method is a non-linear model based

on a smaller neural network, using dense vectors as inputs (e.g., word embeddings) [Goldberg 2016].

Ruder *et al.* [2016b] applied a convolutional neural network (CNN) model for large-scale tasks of authorship attribution based on the work reported [Kim 2014]. They evaluated their method on various text data, such as emails, reviews, blogs, and tweets to obtain excellent accuracy results between 85.0% and 95.0% across the datasets used. Three stylometric feature such as lexical, syntactic, and character-level features were combined [Ding *et al.* 2017] using the work reprorted [Kim 2014] for authorship identification using the international conference on weblogs and social media (ICWSM) labeled Twitter datasets [Al Zamal *et al.* 2012] to achieve the best accuracy of 96.3%.

Ding *et al.* [2017] explored an unsupervised method for extracting stylometric features by integrating various linguistic types into the distribution representation of words to understand the writing style specifically for authorship attribution. The suggested model focuses on capturing the variation in writing—not the actual writing style of the author, on determining the linguistic characteristics in the texts. The features used in the model include lexical, topic, character, and syntactic information extracted from unlabeled texts of the target domain and fed into a single-layer with two output layers for authorship verification and identification problems using PAN2014[1] dataset in English and ICWSM2012[2] Twitter dataset [Al Zamal *et al.* 2012]. They measure the model's performance using the area under the receiver operating characteristic curve (AUROC)–an evaluation measure for binary classifiers where positive and negative labels are equally important, reaching 77% word2vec with skip-grams. They reached a higher accuracy of 85% with lexical and topical features but only 64% with syntactic representation for the identification problem.

Shrestha *et al.* [2017] used CNN, based on a sequence of character n-grams for the authorship attribution problem of posts on social media. Besides indicating a satisfactory result of 76.1% of accuracy for 50 authors based on the dataset provided [Schwartz *et al.* 2013b]. They noticed that 30.0% of the users presented in the dataset behaved as automated bots, turning the task into a straightforward method; however, when the profiles were removed, the classification accuracy dropped to 68.3% for 35 authors. Sari *et al.* [2017] used continuous representation through a neural network combined with a classification layer for authorship attribution. They reached an accuracy of 94.8% on the IMDb62 dataset [Seroussi *et al.* 2010].

Liu and Guo [2019] presented a combination of bidirectional long short-term memory (BLSTM) with CNN layers. They exploited the importance of the features using the attention mechanism for sentiment analysis on the Twitter dataset. They ignored long dependencies that could capture features (or words) of diverse importance. The author in [86] presented a combination of CNN containing max pooling and recurrent neural network (RNN) for valence arousal (VA) prediction on the bases of news posts on social media. They attempted various varieties of CNN with LSTM and gated recurrent unit (GRU) modules on three other datasets of posts on social media. The model was evaluated using the root-mean-square error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient, achieving 1.34%, 0.98%, and 0.77%, respectively.

Zhang *et al.* [2018c] used a new method to encode the syntax parse tree of the sentence into a learnable distributed representation. Mainly, they constructed an embedding vector for each word in the sentence, encoding the path in the syntax tree corresponding to the word. Hu *et al.* [2020a] employed a novel embedding-based framework (DeepStyle) to learn

1 https://pan.webis.de/clef14/pan14-web/
2 https://www.icwsm.org/2012/

the representations of the writing styles of authors of Twitter and Weibo posts. They achieved 59.20% and 58.90% accuracy on Twitter and Weibo with 100 posts from each user for training and testing on 20 posts.

Suman *et al.* [2021] used a capsule network-based model over a character-level n-grams feature to extract features of tweets for author attribution. They performed experiments on the models on the Twitter dataset for 50 authors with the best 68% accuracy. It is challenging or impossible to collect exhaustive or even representative samples of most previous works. It might, therefore, be better to use a model to learn the author's writing. Only one author's training data to extract stylistic features and generate a better writing style could advance authorship attribution performance. Choosing a suitable model to learn the writing style representations from the unstructured data distribution is important for AA tasks. For authorship identification, Jafariakinabad and Hua [2019] applied a syntactic recurrent neural network to encode the syntactic patterns of a document in a hierarchical structure.

Theophilo *et al.* [2021] considered two subsets of a dataset of tweets, exploiting CNN to determine the author. They considered character 4-gram to map information signal in the input text, feeding it to a CNN model to search for a manifold suitable for the author identification in contrast to the previous works [Rocha *et al.* 2016]. They achieved an accuracy of 74.15% with 1000 messages per author compared to 59.36% and 37.08% [Rocha *et al.* 2016; Ding *et al.* 2017].

Table 12 compares related studies based on common characteristics, including the source as the dataset used in the proposed methods, the language of the dataset, the amount of user the dataset contains, the feature and prediction methods, the results, and the question of whether the features can be interpreted or visualised.

In Chapter 6, we focus on exploring the pretraining language model with regularized deep neural network (RDNN) model for PAA systems, which is hard to solve because of the inherent characteristics of the input text. The input texts are casual snippets with short contents and rife with obfuscating vocabularies, such as idioms, onomatopoeias, homophones, phonemes, synonyms, acronyms, anaphora, and polysemy; however, none of the limited earlier studies on PAA focused on interpretability – a critical factor for real-life investigation scenarios.

Chapter 6 further discusses the knowledge divergence by adopting a method that used the extracted features to visualize the results of the authorship attribution. Multiple studies strikingly reported a higher authorship attribution accuracy with several style markers, such as lexical, syntactic, idiosyncratic, and structural markers [Rocha *et al.* 2016; Sari *et al.* 2017; Suman *et al.* 2021; Theophilo *et al.* 2021]; however, the efficiencies of most of the past methods decreased with an increasing number of authors [Sarwar *et al.* 2018a]. They significantly depend on extensive feature engineering to reflect the content and style of an author [Ding *et al.* 2017]. Aslo, lacked transparency and interpretation capability [Wu *et al.* 2021b], essential for language-based pretrial investigation.

The attention-based BLSTM and bidirectional gate recurrent unit (BGRU) incorporated with convolutional max pooling is introduced in this paper for PAA to achieve a better classification performance that could generate a stable and understandable explanation. The ABC methods promise to construct stylometric features by combining two pre-trained language models to learn stylometric features into a distributed representation of words and stacking them into two independent sequential models (BLSTM and BGRU) to translate the extracted text in reverse order and affirm the state of the past context of the text.

The attention mechanism was applied to the output of the BLSTM and BGRU layers of ABC to emphasize style variations allowing the seamless flow of unimpeded information across the network. This was conducted before using convolution and max pooling to reduce features

Table 12: Comparison of related studies based on common characteristics

References	Source	Language	User	Feature	Method	Reesults	Visualize
Ding *et al.* [2017]	PAN2014 ICWSM	English	17K post 1K users	Word embedding	Shallow NN SoftMax	59.4%	No
Hu *et al.* [2020a]	Twitter Weibo	English & Chinese	7K & 9K 1K users	Char, word, bigram POS embedding	Multi-level, CNN Fusion, Triplet loss	51.5%	No
Jafariakinabad and Hua [2019]	Blog IMDb	English	19.3K post 7250 users	Word & POS tags	RNN & SoftMax	92.15%	No
Koppel *et al.* [2011]	Blog	English	19.3K post 7250 users	Character n-grams	Similarity	93.2%	No
Neal *et al.* [2018]	Twitter	English	40/120-200 tweets/user	Stylometric n-gram	Cosine, Euclidean and Manhattan	98.5%	No
Neal *et al.* [2017b]	Blog	English	1000 with 4 post/user	Stylometric n-gram	Isolation Forest	98.0%	No
Rocha *et al.* [2016]	Twitter	English	10M post 10K users	Character n-grams, affix punctuation	SVM RF	59.4%	No
Ruder *et al.* [2016b]	Emails Twitter	English	10-62 users	Character & word embedding	CNN SoftMax	59.4%	No
Sari *et al.* [2017]	News article IMDb	English	10-62 users	Character, word FastText embedding	Shallow NN SoftMax	83.7%	No
Sapkota *et al.* [2015]	News article IMDb	English	10-62 users	Character n-grams, affix punctuation	SVM	69.3%	No
Schwartz *et al.* [2013b]	Twitter	English	1K-2K post 10K users	Character, word n-grams K-signature	SVM	69.7%	No
Seroussi *et al.* [2010; 2011; 2014]	IMDb	English	17K post users	LDA	Hellinger Distance	72.0%	No
Shrestha *et al.* [2017]	Twitter	English	50 users	Character n-grams	CNN, max-pooling & SoftMax	59.4%	No
Suman *et al.* [2021]	Twitter	English	120K 10-50 users	Word level n-gram GloVe, BERT	Capsule network SoftMax	68.0%	No
Theophilo *et al.* [2021]	Research articles	English	80K post 808 users	POS tagging pre-trained embedding	CNN, max-pooling SoftMax	76.9%	No
Wen and Li [2018]	Twitter[3] Yelp[4]	English & Chinese	11K & 2K	Word embedding	CNN, max-pooling LSTM	55.7%	No
Yang and Chow [2014a]	Blog	English	19.3K with 35 post 7250/user	Content-based & style-based	Logistics Regression	80.0%	No
Zhang *et al.* [2018c]	Blog IMDb	English	19.3K post 7250 users	Syntax Parse Tree & Word	CNN	56.70%	No
ABC	IMDb62 Blog50 Twitter50	English	62 users 678K users 1K users	POS tagging, pre-trained embedding	BLSTM, BGRU Attention,CNN, pooling & SoftMax	85.40%	**Yes**

and learn the position-invariant of the author writing for a questioned text in the classification layer. The ABC method was assessed on four datasets of text snippets to investigate its generalization capability across varying input modalities.

6.3 PRELIMINARIES

The principal objective in Chapter 6 is to build an attention-based bidirectional LSTM and GRU integrated with convolution and max pooling mechanisms to improve the accuracy of a PAA system. Figure 39 illustrates the implementation of the proposed ABC model to advance the research on authorship-based tasks. The methodological framework in Chapter 6 includes: *A*. the input layer with the collection of social media data and other related text; *B*. word embedding layer comprising pre-trained word2vec and GloVe to encode the stylistic elements in the text; *C*. bidirectional BLSTM/BGRU, attention, and convolution with max pooling; and lastly *D*. the classification layer, which the actual prediction of the author is made based on the learn writing style of the author. The purpose of agglutinating the embedding layer, the BLSTM-BGRU layer with convolution, and the max pooling layer were to achieve a PAA system with better classification performance, a much faster convergence speed, and high-precision visualization capability.

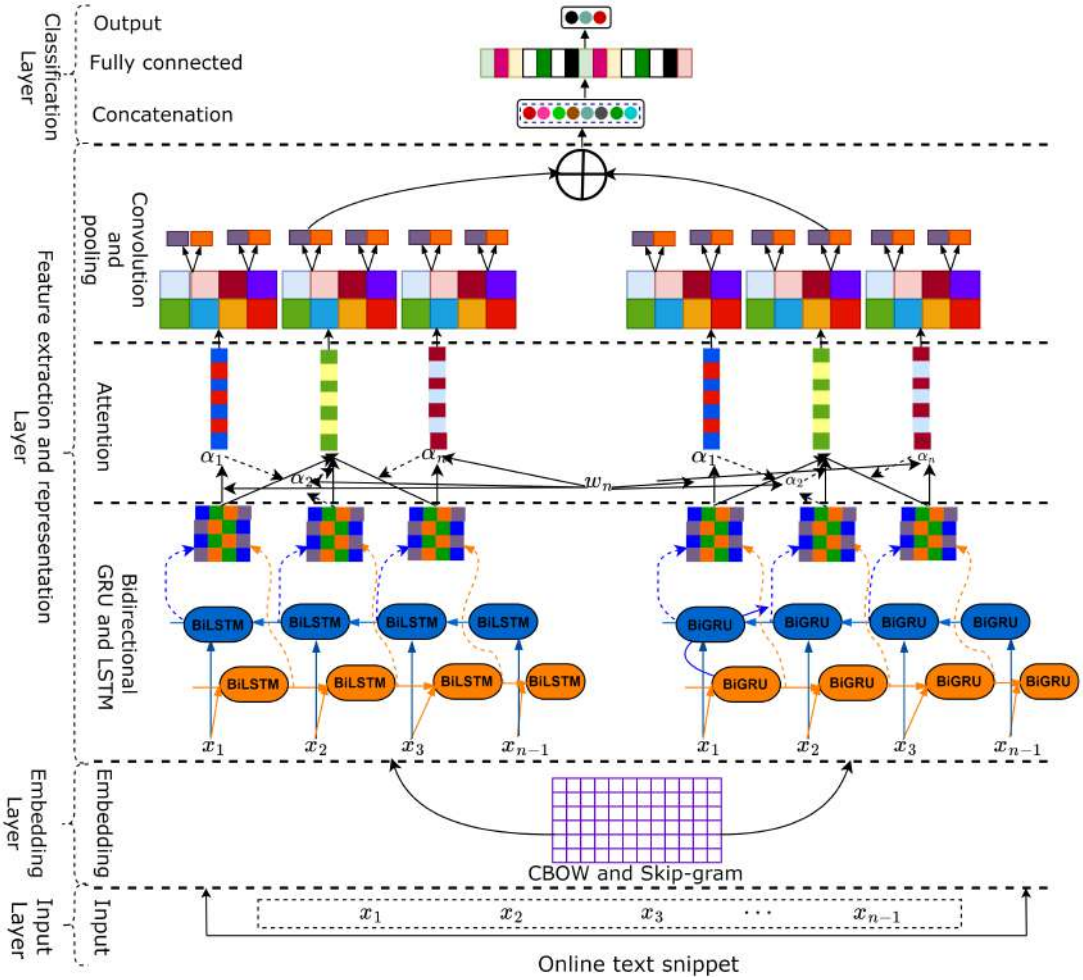


Figure 39: ABC model

6.3.1 *Input layer*

The input layer comprises online text snippets such as movie reviews (like the IMDb62 dataset of prolific writers), Blogs50, and the recently made Twitter datasets shown in Chapter 3.

6.3.2 *Embedding layer*

Natural language processing systems traditionally represent each word as a one-hot vector–a vector filled with 0s except with 1 at the position associated with the word. The one-hot representation is high-dimensional and sparse. Earlier studies for the PAA system failed to capture complete stylometric information, such as the features' lexical, syntactic, structural, and semantic relatedness. Representing words in a continuous vector space has, therefore, been proven efficacious in computational linguistic and NLP tasks by accumulating similar words into a cluster.

Recently, using NN to obtain word representations is adorable and exhilarating because the learner vectors explicitly encode several linguistic and uneven patterns in the texts. The word2vec with Skip-gram [Mikolov *et al.* 2013a] and continuous bag-of-words (CBOW) [Wang *et al.* 2017c] models are two ways of learning word embedding representations. These two models are similar, except that the learning objective of the Skip-gram model attempts to learn word representations by maximizing the probability of predicting surrounding features based

on the middle word. The CBOW finds word vector representations useful for predicting the middle words within a context as shown in Figure 40.

Here, the derivation of the CBOW model is presented as the CBOW model, and the Skip-gram model are mirror images of each other. Owing to the high computational complexity of the original CBOW model and the Skip-gram model, hierarchical softmax or negative sampling was applied in the training processes. Hierarchical softmax uses a frequency-based Huffman tree, such as a binary tree, to represent all the words in the vocabulary at the output, which are leaf units [Hu *et al.* 2016a]. In the CBOW model with hierarchical Softmax, the output layer is replaced by a Huffman tree. The hidden layer averages the input word vectors; therefore, the output of the hidden layer is:

$$h = \frac{1}{C} \sum_{u \in (\text{context})(w)} v(u) \tag{28}$$

where $v(u)$ denotes the vector of the word u, context(w) is the set of contextual information's of the word w, and C is the cardinality of the set context(w). Therefore, in each context, the conditional probability of the word w can be defined as:

$$p(w|\text{context}(w)) = \prod_{j=1}^{k(w)-1} \|h^T v'_{n_{(w,j)}}\| \tag{29}$$

where $n_{(w,j)}$ represent the j-th inner point from root to word w in the Huffman tree $v'_{\eta_{(w,j)}}$ is the vector of an inner point n, where $k(w)-1$ represents the length of the Huffman tree $\mathcal{H}$ for word w and $\|\|$ is a function defined as:

$$\|x\| = \sigma(x)^{d^w_{j+1}} [1 - \sigma(x)^{(1-d^w_{j+1})}] \tag{30}$$

where d^w_{j+1} is the j-th bit of the $\mathcal{H}$ for word w. Thus, we implemented it by maximizing the conditional probability in Equation 29 during the training of the model shown in Figure 39 for the context (or target) of the word w. The log of the conditional probability provides the loss function as:

$$l = \log p(w|\text{context}(w)) \tag{31}$$

The derivative was obtained l as a loss function regarding the vector of the inner point $\eta_{(w,j)}$ as follows:

$$\frac{\partial l}{\partial v'_{n_{(w,j)}}} = \frac{\partial l}{\partial h^T v'_{n_{(w,j)}}} \frac{\partial h^T v'_{n_{(w,j)}}}{\partial v'_{n_{(w,j)}}} = h^T \|1 - h^T v'_{n_{(w,j)}}\| \tag{32}$$

where $j = 1, 2, \cdots, l(w)-1$ we define the derivative of l regarding vector of information contextual of words u as:

$$\frac{\partial l}{\partial v(u)} = \sum_{j=1}^{l(w)-1} \|1 - h^T v'_{n_{(w,j)}}\| v'_{n_{(w,j)}} \tag{33}$$

The two are mirror images of each other. The learning objective of the CBOW model is to learn a word vector to predict the centered word within a particular context; the Skip-gram is used to learn a word vector to predict surrounding words based on the centered word. The word representation, such as word vectors of the input text from the input layer is learned by maximizing the l function by applying the variational stochastic gradient descent method [Shen *et al.* 2016], summarized in the Word2Vec–CBOW training Zeile 4.11.

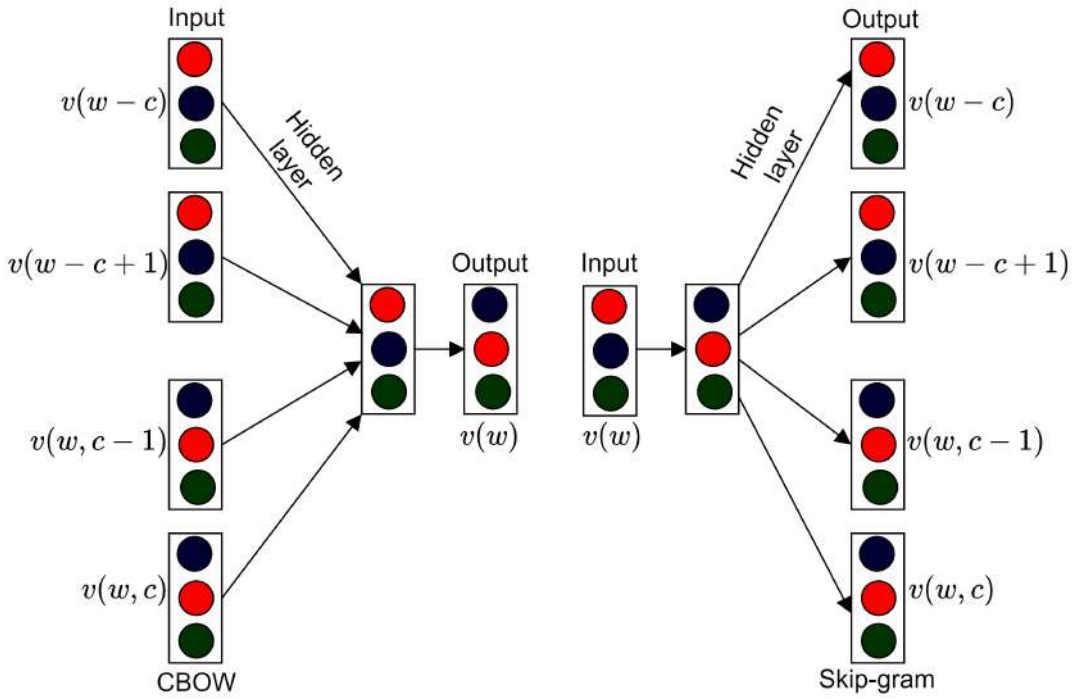


Figure 40: CBOW and the Skip-gram model for word embedding representations.

Algorithmus 4 : Word2Vec–CBOW Pseudocode

input : Initialize the Huffman Tree $\mathcal{H}_t$, n as learning rate, ϕ as the maximum training epoch, and the dimension of the word embedding as D.

output : v is the word embedding

4.1 initialize parameters;
4.2 initialize vectors of words in the input text: v and vectors of inner points of $\mathcal{H}_t : v'$;
4.3 **for** $i = 1$ *to* ϕ **do**
4.4 update vectors of inner points;
4.5 **for** $j = 1$ *to* $l_{(w)} - 1$ **do**
4.6 $g = \|1 - h^T v'_{n_{(w,j)}}\|$;
4.7 $e = e + g v'_{n_{(w,j)}}$;
4.8 $v'_{n_{(w,j)}} = v'_{n_{(w,j)}} + n \times h^T g$
4.9 update vectors of contextual information of words points;
4.10 **for** *u in Context(w)* **do**
4.11 $v(u) = v(u) + \eta \times e$

6.3.3 *Feature extraction and representation layer*

The feature extraction and representation layers comprise CNN, LSTM, bidirectional LSTM, and GRU to model the author's writing style for PAA systems. CNN, LSTM, and LSTM-CNN models have been used for AA tasks with the same dataset [Ruder *et al.* 2016b; Sari *et al.* 2017; Zhang *et al.* 2018c]. This section briefly describes the building block procedure.

6.3.3.1 *Bidirectional long short-term memory and gated recurrent unit layer*

An LSTM is a variation of RNN architecture designed to solve the vanishing gradient. LSTM generates its output based on the input from the time step and the outcome of the prior time step; it sends the output to the next time step, such as RNN, but with no memory. Each LSTM unit comprises a memory cell preserving its state over arbitrary time intervals and three non-linear gate, including an input gate$_{i_t}$, a forget gate$_{f_t}$, and an output gate$_{o_t}$. These gates regulate the information flow in and out of the memory cell. If σ, $\tanh(\bullet)$, and $\odot$ are the element-wise sigmoid function, hyperbolic tangent function, and the product operation, respectively.

In LSTM, x_t and h_t are the input vectors and the hidden state vectors at the time t. Similarly, μ and w are the weight matrices of the updated gates or cells for the input x_t and hidden state h_t, and the variable b_t represents the bias vector of the update gate. The forget gate decides which information needs to be forgotten by outputting the author's writing style:

$$f_t = \sigma(\mu_f x_t + w_t h_{t-1} + b_f) \tag{34}$$

and the input gate decides what new information should be kept by computing i_t and $\vec{c}_t$, therefore, the combination of the two resulted into:

$$\begin{aligned} i_t &= \sigma(\mu_i x_t + w_i h_{t-1} + b_i) \\ \vec{c}_t &= \tanh(\mu_c x_t + w_c \times h_{t-1} + b_c) \\ c_t &= f \odot c_{t-1} + i_i \odot \vec{c}_t \end{aligned} \tag{35}$$

To capture the future context besides the preceding context, the hidden state of the forward BLSTM $\overrightarrow{h}_t$ was combined with the backward $\overleftarrow{h}_t$ hidden layers. This resulted in a temporal information flow in both directions to learn the lexical and stylistic features based on the texts for post-authorship analysis. A BGRU is a simpler variant of LSTM with two gates, an update gate r that combines forget and input gates, and a reset gate z [Liu and Guo 2019; Li *et al.* 2020]. Like LSTM, the update and reset gate is calculated as follows:

$$\begin{aligned} r_t &= \delta(u_r x_t + w_r \times h_{t-1} + b_r) \\ z_t &= \delta(u_z x_t + w_z \times h_{t-1} + b_z) \end{aligned} \tag{36}$$

where $\delta(\cdot)$ is the logistic sigmoid function, u, (u, w) and b remain the same. The reset gate decides when the previous hidden state should be ignored, and the update gate decides the information that should be passed to the current state. In a BGRU layer, the t-th hidden h_t with the reset gate r_t and update gate z_t is computed as follows:

$$\begin{aligned} h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \\ \tilde{h}_t &= \tanh(u_{\tilde{h}_t} x_t + w_{\tilde{h}_t}(\tilde{h}_{t-1} \odot r_t)) \end{aligned} \tag{37}$$

6.3.3.2 *Attention layer*

Attention models are used to assign various weights to words contributing differently to the writing style of BLSTM and BGRU. A customary practice of assigning various weights to diverse writing styles of an author in an unlabeled text is to use a weighted combination (See Figure 41) of all hidden states, X_{A_w} as follows:

$$\begin{aligned} \alpha_t &= \frac{e^{(v^T \times \tilde{h})}}{e^{(v \times \tilde{h})}} \\ X_{A_w} &= \sum_t (\alpha_t h_t) \end{aligned} \tag{38}$$

where $\tilde{h}_t$ and h are defined as revealed in Equation 38 and is a trainable parameter [Li *et al.* 2020; Zhou *et al.* 2021].

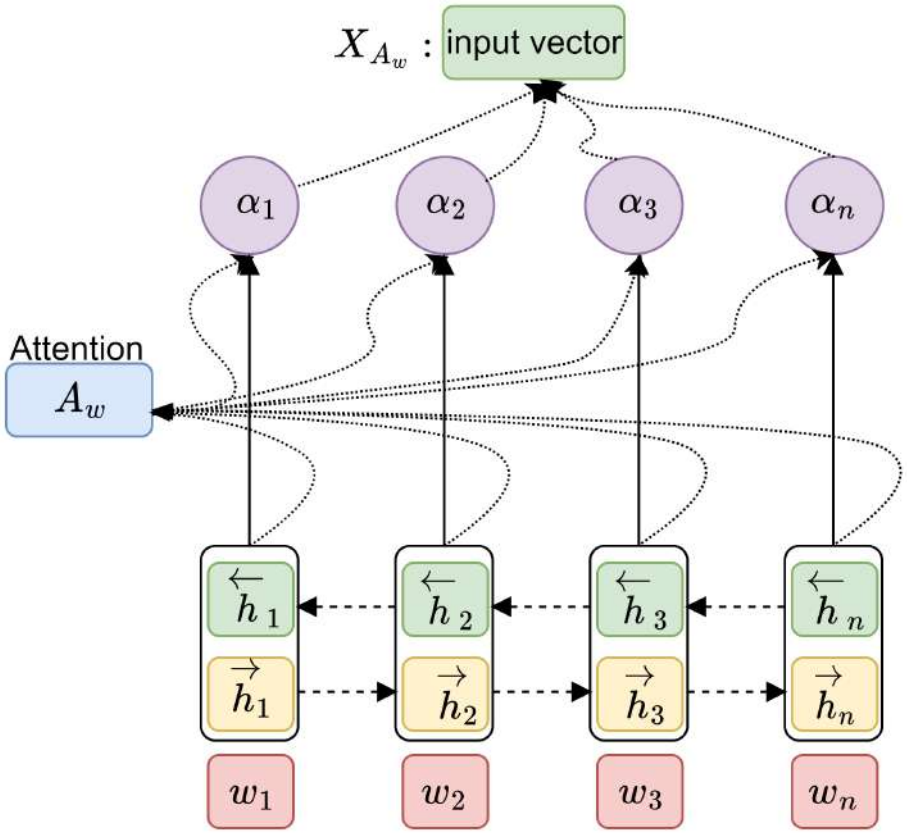


Figure 41: The attention mechanisms are fitted with bidirectional LSTM and GRU networks.

6.3.3.3 *Convolutional neural network layers*

Convolutional neural networks (CNNs) comprise several convolution layers. These layers are used in NLP applications to better understand the local distinctive feature. The study conducted convolution operations on the feature vector from the attention layer by adding a linear filter. For a provided post on social media in a sentence X with distinct x words, first, the embedding vector of size e was created, then a filter F of size $e \times h$ was used repeatedly as a sub-matrices to represent the input data. The results of this generate a feature map $M = [m_1, m_2, \cdots, m_{x-h}]$ as follows:

$$m_i = F \times X_{i:i+h-1} \tag{39}$$

where $i = 0, 1, \cdots, x-h$ and $X_{i:j}$ is a sub-matrix of X from row i to j as a common practice to the dimension of feature maps by feeding them to a pooling or sub-sample layer. The max pooling is a common pooling strategy, selecting the most important feature b of the map as follows:

$$b = \max_{0 \leqslant i \leqslant x-h} (m_i) \tag{40}$$

The results of the pooling layer are concatenated to form a pooled feature vector used as the input to the fully connected layer (FCL) (See Figure 42).

6.3.3.4 *Fully connected layer (FCL)*

In an FCL, the feature vector representation derived from the weight vector of the concatenated pooling layers is mapped to the input vector through a weights matrix to learn an author's

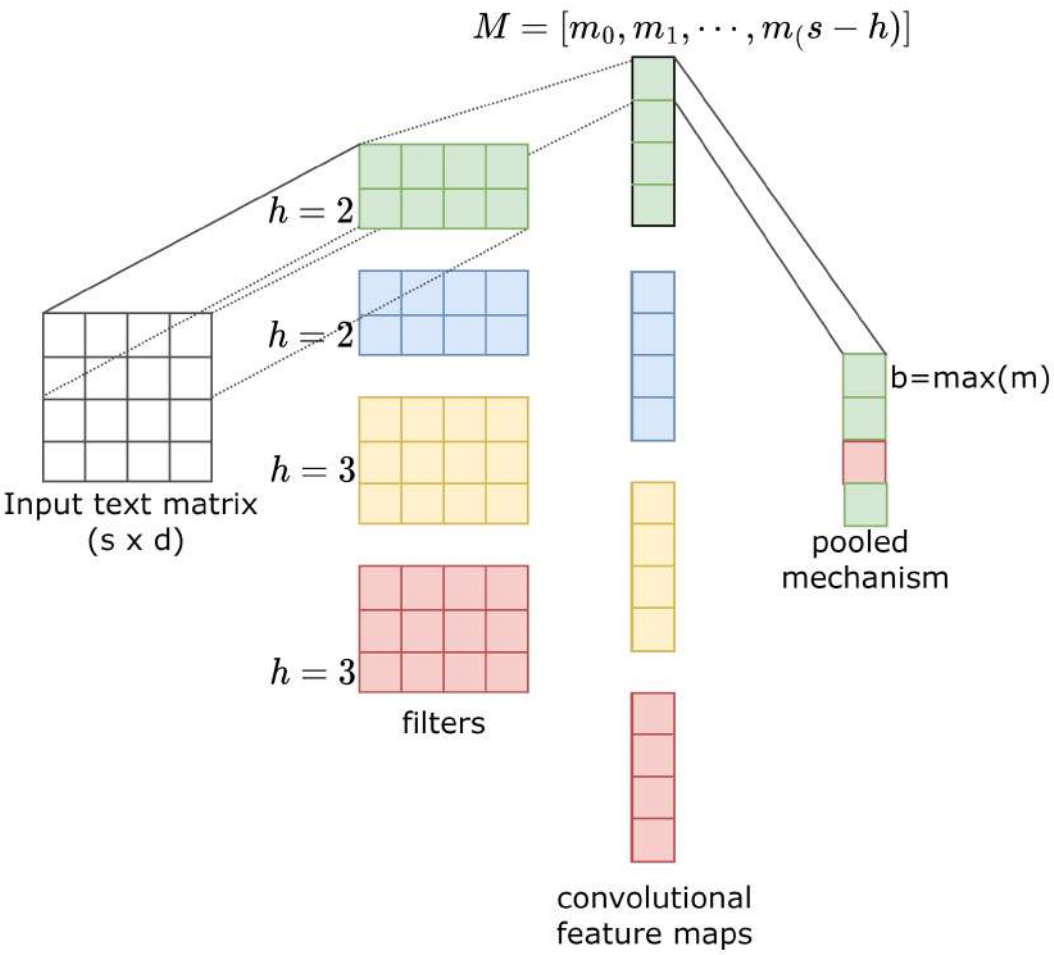


Figure 42: Convolutional layer to extract local feature.

writing style for ABC model in Figure 39. The FCL includes multiple dense layers, non-linear activation, softmax, and prediction function to obtain the correct post-classification as follows:

$$\mathbb{H}_t = \text{SoftMax}(w_t h_{t-1} + b_t) \tag{41}$$

where w_t and b_t are parameter learned in training, $\mathbb{H}_t$ is the obtained from the pooled concatenated feature vector and h_{t-1} is the feature map received from the CNN layers. The output layer is used to perform the correct classification using the softmax function as in Figure 39. To learn the model parameters, the cross-entropy loss was minimized as the training objective using the Adam optimization algorithm [Kingma and Ba 2014]. It is provided by

$$\text{CrossEntropy}(p,q) = -\sum_{p}(\text{x})\log(q(x)) \tag{42}$$

where p is the true distribution of a one-hot vector representing characters in the messages posted on social media, and q is the output of the softmax. This corresponds to computing the negative log probability of the true authors.

6.4 PROPOSED SYSTEM

The limitations of deep learning architectures to improve the authorship attribution system are discussed. A bidirectional GRU was developed with an attention-based CNN (ABC)

integrated with pre-trained word embedding methods, such as word2vec CBOW and GloVe, converting words in the message post on social media into meaningful vectors. In expansion, a bidirectional LSTM, bidirectional GRU, and attention mechanism captured the long-range dependencies information, while the CNN layer obtained the local writing style of the author for the unquestioned online text snippets disregarded [Sari *et al.* 2017; Suman *et al.* 2021].

Thus, the input and pre-trained embedding matrix $w_g \in \mathbb{R}^{(n \times e)}$ where n and e denoted the total the total of words and embedding dimension by combining word2vec CBOW and GloVe were generated to embed a comment vector $c \in \mathbb{R}^m$ with m being the padding length or the maximum number of words $w_t t \in [1, m]$ represented in the comment or message posted on the social media as follows:

$$c_t = w_g w_t, t \in [1, m] \tag{43}$$

The two parallel layers of BLSTM and BGRU were applied to the output of the embedding layer to process the sequences' arbitrary lengths of post messages on social media. Dependencies were extracted in forward and backward directions through LSTM (e. g. short and longer sequences) as follows:

$$\begin{aligned}
\overrightarrow{h}_{t_{\text{BLSTM}}} &= \overrightarrow{\text{BLSTM}}(c_t), \mathrm{t} \in [1 - m] \\
\overleftarrow{h}_{t_{\text{BLSTM}}} &= \overleftarrow{\text{BLSTM}}(c_t), \mathrm{t} \in [\mathrm{m} - 1] \\
\overrightarrow{h}_{t_{\text{BGRU}}} &= \overrightarrow{\text{BGRU}}(c_t), \mathrm{t} \in [1 - m] \\
\overleftarrow{h}_{t_{\text{BGRU}}} &= \overleftarrow{\text{BGRU}}(c_t), t \in [m - 1]
\end{aligned} \tag{44}$$

For each word w_t, a structural feature vector was obtained by concatenating forward and backward contexts as follows:

$$\begin{aligned}
h_{t_{\text{BLSTM}}} &= [\overrightarrow{h}_{t_{\text{BLSTM}}}, \overrightarrow{h}_{t_{\text{BLSTM}}}] \\
\mathrm{h}_{t_{\text{BGRU}}} &= [\overleftarrow{h}_{t_{\text{BGRU}}}, \overleftarrow{h}_{t_{\text{BGRU}}}]
\end{aligned} \tag{45}$$

The attention mechanism $\mathrm{h}_{t_{\text{BLSTM}}}$ and $\mathrm{h}_{t_{\text{BGRU}}}$ as revealed in 45 was then applied to enable the proposed ABC model to pay more or less attention to diverse linguistic modalities for the message posted on social media to learn the writing style of each author. To achieve this, feature vector was adjusted by extracting informative semantic words in the input online snippets:

$$\begin{aligned}
u_{t_{\text{BLSTM}}} &= \tanh\left(w_{t_{\text{BLSTM}}} h_{t_{\text{BLSTM}}} + b_{t_{\text{BLSTM}}}\right) \\
u_{t_{\text{BGRU}}} &= \tanh\left(w_{t_{\text{BGRU}}} h_{t_{\text{BGRU}}} + b_{t_{\text{BGRU}}}\right) \\
\alpha_{t_{\text{BLSTM}}} &= \frac{e^{\left(u_{t_{\text{BLSTM}}}^T \times u_{t_{\text{BLSTM}}}\right)}}{\sum_t e^{\left(u_{t_{\text{BLSTM}}}^T \times u_{t_{\text{BLSTM}}}\right)}} \\
\alpha_{t_{\text{BGRU}}} &= \frac{e^{\left(u_{t_{\text{BGRU}}}^T \times u_{t_{\text{BGRU}}}\right)}}{\sum_t e^{\left(u_{t_{\text{BGRU}}}^T \times u_{t_{\text{BGRU}}}\right)}} \\
X_{t_{\text{BLSTM}}} &= \sum_t \alpha_{t_{\text{BLSTM}}} \times h_{t_{\text{BLSTM}}} \\
X_{t_{\text{BGRU}}} &= \sum_t \alpha_{t_{\text{BGRU}}} \times h_{t_{\text{BGRU}}}
\end{aligned} \tag{46}$$

where u_t denote the hidden representation of h_t and u_w is the context vector initialized randomly and jointly learned during the training phase. The importance of a word u_t is measured using cosine similarity between u_t and u_w normalized as indicated by $\alpha_{t_{\text{BLSTM}}}$ and $\alpha_{t_{\text{BGRU}}}$ in Equation 46. The importance weights α_t are aggregated into x using the weighted sum, where x denote the feature vector and summarization all the information of the author's

writing style in the data. After obtaining the feature vector denoted as x a convolution operation was used to extract informative features to reduce dimensionality in the input text. The convolution operation enables the proposed ABC model to find the positional invariants in the author's writing style for an unlabeled post.

In the ABC model, two *1*-D CNNs were employed with a fixed number of filters and various window sizes for each branch (i.e., the outputs of BLSTM and BGRU units) independently. Four outputs from the CNN layer were observed with two independent CNN applied to capture the long-term dependencies from BLSTM and BGRU layers. The output of the CNN layer was stacked independently to the maximum and average pooling mechanism to downsample the feature maps. This ensured more robust resulting feature vectors to find positional changes in the author's writing style. For example, by denoting the number of filters in the CNN layer f then the final feature vector l_c for each pooling operation is $l_{ci} = l_{c1}, l_{c2}, \cdots, l_{cf}$, where i is an integer number from $1, 2, \cdots, 8$. These feature vectors are concatenated to form the conclusive feature vector that could be learned by an author's writing style, based on anonymous messages posted on social media to improve the PAA system.

The observed $l_c = [l_{c1}, l_{c2}, \cdots, l_{c8}]$ represent the obtained feature vector representation of distributed word representation of each author and then standardized by applying batch normalization to accelerate ABC model during training to reduce overfitting and generalize for unlabelled online text snippiest or documents. Finally, an FCL was used to transform $\mathrm{l_c}$ into a high-level feature vector representation, which enabled the proposed ABC model to learn the author's writing style representation. The output of this layer is calculated using Equation 41.

6.5 EVALUATION AND RESULTS

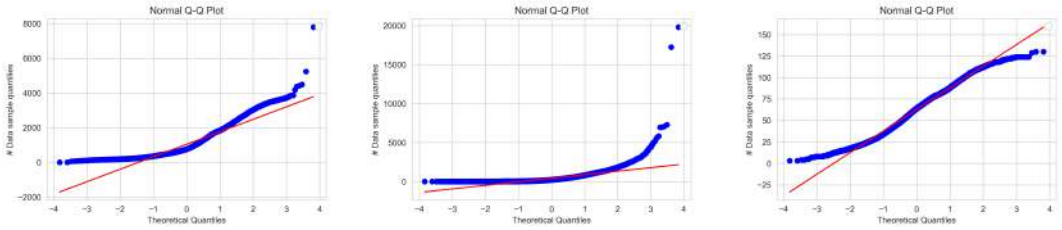
In this section, we evaluates the performance of the proposed ABC on a large-scale social media and publicly available dataset in English text as discussed in Chapter 3. The experiments were implemented using Ubuntu 20.04, Python 3.9 with Keras TensorFlow 2.4.2 on AMD Ryzen 9 3900XT 12-Core Processor GPU RTX 2080 Ti 64 GB RAM.

6.5.1 *Parameter settings*

The Keras library in the TensorFlow machine learning framework was used to implement the proposed ABC model for the PAA system and the other baseline models. The objective was to minimize the complexity of the model by removing unnecessary elements, such as the number of hidden nodes, and in the dense layer by finding optimal hyper-parameters with the hyper-band method in Keras tuner [Chowdhury *et al.* 2022]. An input matrix of 50,000 words was constructed to divide the raw input data into tokens, which helped the ABC model to understand the context and interpret the vital information in the text by analyzing the word sequence using tokenization in the Keras library. A preprocessing step was applied before tokenization by removing irregular text formats, text content loss, and incomplete and duplicate documents.

Words in the text adding no meaning to the sentence were ceased; they would not affect text processing for the defined purpose and were removed from the vocabulary to reduce noise and the dimension of the feature set. In order to demonstrate the statistics and the distribution of the word frequency in each dataset, we plot a theoretical quantiles (e. g. Q-Q (i. e. quantile-quantile)) as the standard normal distribution to establish if the x-axis (e. g. each dataset) follows a Gaussian distribution or not on the y-axis. As indicated in Figure 43 (a) the points plotted on the graph are not perfectly positioned on a straight line but are significantly scattered from the positions; this demonstrates that the distribution of each candidate author's writing style did not align with the standard normal variate. While, in Figure 43 (b), the two points at the ends of the curve are demolished from the straight line as opposed to one

from the beginning, which signifies that the writing styles of the authors in the texts are not normally distributed. In Figure 43 (c), the two points from the curve fell out of the straight line as against one from the beginning, which demonstrates that a relationship between the dataset (e. g. x-axis and the normal theoretical quantiles on the y-axes signifies that the writing style of each author is not normally distributed (e. g. deviation from the normal distribution). Generally, Figure 43 revealed the first 100 words of online text snippets for movie reviews and blogs were considered, whereas the first 45 words were considered with tweet datasets by setting the padding length to 100 and 45, respectively.



(a) IMDb62 (b) Blog (c) Twitter

Figure 43: Quantile-quantile statistical distribution of the datasets

Publicly available pre-trained GloVe vectors and word2vec with CBOW concatenated formed the weights in the embedding layer. The 100-dimension of word2vec was used, trained on word vectors of three million words and phrases of roughly 100 billion words from a Google News dataset [Mikolov *et al.* 2013a]; similar to GloVe trained on Wikipedia [Pennington *et al.* 2014], it contained 6 billion tokens with a vocabulary size of 400,000.

The embedding size for movie reviewers (IMDb) and blog datasets was 200, with a fixed size of 100 for the tweet datasets. In the BLSTM and BGRU, each neuron spanned between 32 and 256 memory units with a step size of 32, but 416 for the BLSTM and 320 for the BGRU provided an optimum value with the Bayesian optimization in the Keras tuner. The library was used to establish the optimum value while restricting the number of iterations to a low value. The maximum number of trials was set to between five and 10, corresponding to two to three per execution trial with a dropout of 0.4. In the convolutional layer, 480 filters with a kernel size of four and six provide the optimum values (see Figure 39).

The size of the FCL was 416, initializing word embeddings using Glorot uniform initialization [Glorot and Bengio 2010b] for the model to converge over a softmax classifier. The entire model was trained for 20 epochs using the Adam stochastic optimizer [Kingma and Ba 2014]. A mini-batch size of 64 yielded better performance for tweets and 32 for blog and movie review datasets when the class label had over 10 or 20 authors; however, the learning rate of 0.0001 and the dropout of 0.75 varied from 0.3 to 0.7 and were constant for all the training datasets, irrespective of the class label. The softmax function was employed in the output layer without the hashing trick.

Finally, the training process was accelerated on the dataset with a class label of less than 50 by setting the learning rate, embedding size, mini-batch size, and the number of epochs to 0.001, 50, 32, and 20, respectively. A 5-fold cross-validation and early stopping with monitoring validation loss in max mode with the patience of five trials were applied in the training process to prevent overfitting.

6.5.2 *Evaluation criteria*

To evaluate the performance of the proposed ABC, we used four statistical metrics such as precision (p_r), recall (r_c), F-measure (f_m), and accuracy (acc). These metrics are mostly used in text classification, information extraction, NLP, sentiment analysis bioinformatics, and authorship

attribution system to identify the classes of an input text correctly and to migrate the degree of missed-classification in the model [Sokolova *et al.* 2006]. All the metrics are class-specific, represent a certain class of author with the label (l) true positive (t_p) represents the number of online text snippets from the dataset with the label (l), false positive (f_p) represents the number of samples with any other author or labels except (l) assigned incorrectly with the label (l), and false negative (f_n) represents the number of samples with the label (l) incorrectly assigned with any other label exception (l). Precision (p_r) was used to measure the exactness of the proposed model as a classifier and recall (r_c) represents its effectiveness. The tradeoff between precision (p_r) and recall (r_c), defines F-measure (f_m) as the weighted harmonic mean of precision and recall (e. g. a composite measure that benefits the model with higher sensitivity and challenges algorithms with higher specificity). The F-score f_β is commonly used in evaluating classification application when the datasets are evenly balanced with $\beta = 1$, and when it favours precision (p_r) and recall (r_c) with $\beta > 1$, it is known as F-measure (f_m).

6.5.3 *Results*

The proposed ABC model is quantitatively compared with various stacking methods for AA tasks described in Table 12. 5-fold cross-validation method was applied for a pre-fold run time of most methods, with a 10% randomized split off as a stratification test set. The test was kept constant; therefore, each training set was a validation set, with a seed to reduce randomness by 10% further. The accuracy and F1-score were determined across 20 epochs, with a 0.60.20.2 split between training, validation, and testing as the quantitative performance indicators.

6.5.3.1 *Empirical results*

The quantitative results are based on two of the evaluation criteria, as shown in Table 13, Table 14, and Table 15 for 50, 20, and 10 authors. The ABC model recorded good results for both acc and f_β for two of the datasets when analysed in relation to other SOTA methods. The ABC model concomitantly showed the highest *acc* of 61.85% and 85.40%, respectively, on the blog50 and 62% and 85% for and f_β on the Twitter50 datasets. The ABC model was very superior, with more than 0.66% on the blog dataset when compared with the most common baseline models, e. g.the SRNN model [Jafariakinabad and Hua 2019] and a margin of at least 4% on the Twitter dataset when compared with the most common baseline models, such as the AC-BLSTM [Liu and Guo 2019]. But the ABC model only got 90.65% *acc* and 90.60% f_β on the IMDb62 dataset. It failed to perform as well as the model that learns the author's writing style from an embedding vector through an n-gram character or word, like Syntax-CNN [Zhang *et al.* 2018c], N-gram+Style [Sari *et al.* 2018b], and Continuous n-gram [Sari *et al.* 2017]. Compared to the BertAA model [Fabien *et al.* 2020a] and a capsule-based CNN [Suman *et al.* 2021], the ABC model got the best acc values of 61.85% and 85.40% on the blog50 and twitter50 datasets. The ABC model obtained precision gains of 1.26%, 4.9%, 2.02%, and 5.50%, respectively. The ABC model outperformed all the models that employed pre-trained embedding by an average of 7.05% in acc and 16.0% in the IMDb62 datasets. The accuracy values for the ARC, CRNN, IWV, and AC-BLSTM [Wen and Li 2018; Wang *et al.* 2016a; Rezaeinia *et al.* 2019; Liu and Guo 2019] methods on the Twitter50 dataset were 77.7%, 65.50%, 58.70%, and 74.89%, respectively. The test results show that the suggested model did better than all the models that used pre-trained embedding by an average of 7% in acc and 16% in f_β, and the difference was the same across all datasets. This result demonstrates that the ABC model is more suitable for textual content with subjective elements, which are among the majority of the textual content of online social media and are good for enhancing authorship attribution on social media.

Table 13 shows the acc and f_β with no extra features over 20 epochs on the Blog50 datasets compared to the BertAA [Fabien *et al.* 2020a]. It also shows the performance of several CNN-based architectures, such as character-level CNN [Ruder *et al.* 2016b] and SRNN [Jafariakinabad and Hua 2019]. The ABC model was able to achieve a level of accuracy of 62.45% on the Blog50 dataset, which is 1.26% better than the style-HAN, and 85.40% on

Twitter50, which is 4.9% better than the previous character-level CNN [Ruder *et al.* 2016b]. The primary characterization of the blog corpus is the text, with an average of 100 tokens. The number of texts per author is still relatively high, with a slight standard deviation. This signifies that the proposed model is suitable for datasets with short sentences and large but balanced text quantities per author.

Additional experiments were conducted on the IMDb62 dataset with a larger number of authors, with 1000 sentences per author. In order to replicate the setup of most methods in Table 13, cross-validation of five over 20% of the data was used as a test sample. The split was made randomly since no standardised training and testing corpus exists for the datasets. Table 13 reports the performance of the proposed model compared to various methods, including the current SOTA. When compared to the majority of the other methods, which range between 70.0% and 95.21%, the syntax-enriched CNN reached 96.16% superiority [Zhang *et al.* 2018c]. While IMDb62 has 1000 training samples for each author, the proposed model did not do as well over 20 training cycles as continuous N-gram [Sari *et al.* 2017], which only found out which topics each author liked rather than their writing style, N-gram+Style [Sari *et al.* 2018b], Syntax CNN [Zhang *et al.* 2018c], and BertAA [Fabien *et al.* 2020a]. This emphasises the limitations of the proposed model in a situation with less training data and insufficient data to capture writing styles from more authors.

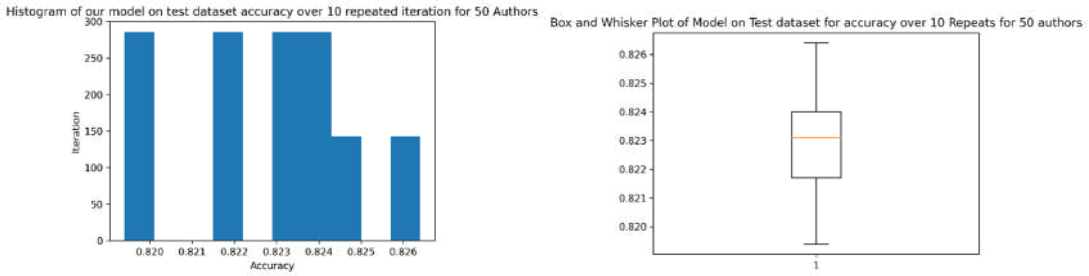
Table 13: Comparison of the results obtained on three datasets for 50 authors

Methods	IMDb62		Blog50		Twitter50	
	acc	f_β	*acc*	f_β	*acc*	f_β
Imposters [Koppel *et al.* 2011]	70.00	76.90	22.60		52.50	
LDAH-S [Seroussi *et al.* 2011]	80.00	72.00	18.30		38.30	
PMSVM [Rocha *et al.* 2016]					59.36	
Stylometric [Alonso-Fernandez *et al.* 2021]					70.60	
N-gram+Style [Sari *et al.* 2018b]	**95.93**					
CNN character [Ruder *et al.* 2016b]		**91.70**		49.40		**86.80**
Lexical & Topical [Ding *et al.* 2017]	95.90		53.10		37.08	
Cont. n-gram[5] [Sari *et al.* 2017]	**94.80**		52.80			
CNN n-gram [Shrestha *et al.* 2017]					56.20	
Syntax-CNN [Zhang *et al.* 2018c]	**96.16**		56.73			
SRNN [Jafariakinabad and Hua 2019]			61.19			
DeepStyle [Hu *et al.* 2020a]	90.51				64.80	
BertAA [Fabien *et al.* 2020a]	93.00		59.70			
Multiple-messages-CNN [Theophilo *et al.* 2021]					74.15	
Capsule-based CNN [Suman *et al.* 2021]					55.40	
ARC [Wen and Li 2018]	77.70		42.65	41.00	59.99	60.00
CNN-LSTM [Wang *et al.* 2016a]	65.50	65.00	47.47	45.00	75.32	75.00
CRNN [Wang *et al.* 2016a]	58.70		33.78	34.00	55.90	55.00
IWV [Rezaeinia *et al.* 2019]	57.40		40.69	40.00	58.60	58.00
AC-BLSTM [Liu and Guo 2019]	74.8	74.00	48.72	45.00	79.98	80.00
Our Approach	90.65	90.60	**61.85**	**62.00**	**85.40**	**85.00**

The effectiveness of the ABC method is further investigated when 10 and 20 authors were randomly selected to perform the experiments on the blog and Twitter datasets as conducted [Ruder *et al.* 2016b; Zhang *et al.* 2018c; Jafariakinabad and Hua 2019]. To analyze the behavior

of the developed system on these datasets with diverse authors, ten sets of 10 and 20 authors were created with 1000 texts for each. These sets were created after randomly choosing authors from the datasets. The study demonstrated that the proposed model has a variance in its prediction. This was conducted by repeating the model and evaluating it on the Twitter50 dataset.

The model was divided into parts and evaluated as functions repeatedly. The evaluation function investigated the effectiveness of the model on the test dataset by fitting the model and the compiler with both training and validation data. First, the ten independently run tests were conducted while totalling the distribution scores concerning the mean values of 82.30% and standard deviation of 0.002%. The results demonstrate that the proposed ABC model adopts a stochastic distribution with unique numerical precision, which is reasonable, as revealed in the histogram of the accuracy scores in Figure 44 (a). Also, in Figure 44 (b), a box and a whisker were plotted to demonstrate a line at the median with 82.20% accuracy on the test set and the interquartile range, or the middle 50% of the samples between about 82.40% and 82.20%. The result demonstrates performing of the proposed model of test scores trained on the same dataset.



(a) Histogram of the accuracy

(b) Box and whisker plot

Figure 44: Performance accuracy and loss curves over epoch on the CCAT50 datasets.

The results in Table 14 and Table 15 demonstrate the performance of the proposed model differs among authors on the blog and Twitter datasets. The highest accuracy of 65.00%, 65.00%, and an F1-score of 64.04%, and 65.00% for blogs 10 and 20 were recorded, respectively. The SOTA method demonstrated a superior capability to capture stylistic information on Twitter 10 and 20 datasets than the current model. The effectiveness of the proposed model increased with various authors and the training data.

As opposed to previous SOTA methods, the efficiency dropped with a variation in the candidate author and training set. These improvements are mainly beneficial by considering the long dependencies in the input text using BLSTM and BGRU layers; assigning various weights to diverse lexical words or phrases in the datasets based on the importance achieved using the attention mechanism; varying the positional invariant of the author's writing style to identify the writer by applying various filter sizes in the CNN layers.

The effectiveness of the ABC method is further presented graphically based on the accuracy and loss curves for visual validation. Figure 45 demonstrates the accuracy and loss curves for the Twitter50 datasets. It can be observed from the figures that the proposed method has better performance in accuracy with faster convergence speed in the training and testing processes on the Twitter50 datasets; however, complex interactions among diverse stylistic features with nonlinearities can be observed in the Twitter50 datasets. Although performance improvement was achieved with the ABC method, the effect of certain features, such as hashtags, was highly indicative.

This distorted the boundaries of the proposed method in understanding the writing style of various authors. The Twitter50 dataset reveals a bias in these ways. The input post was not preprocessed but was converted to lowercase to erode the character sensitivity effect of a discrepancy between upper- and lowercase letters. The reverse domain name notation of

Table 14: Results for 20 authors

Methods	Blog20		Twitter20	
	acc	f_β	acc	f_β
Imposters [Koppel *et al.* 2011]	35.40	35.00	71.40	71.00
LDAH-S [Seroussi *et al.* 2011]	52.50	52.00	90.00	84.00
CNN-word [Ruder *et al.* 2016b]	57.00	57.00	94.20	94.20
CNN-word-char [Ruder *et al.* 2016b]	60.10	60.00	95.50	95.00
Syntax-CNN [Zhang *et al.* 2018c]	64.10	64.10		
BertAA [Fabien *et al.* 2020a]	64.10	64.10		
ARC [Wen and Li 2018]	52.30	52.00	63.15	63.00
CRNN [Wang *et al.* 2016a]	59.21	59.00	72.30	73.00
CRNN [Wang *et al.* 2016a]	59.21	59.00	72.30	73.00
IWV [Rezaeinia *et al.* 2019]	53.60	53.00	71.50	71.00
AC-BLSTM [Liu and Guo 2019]	62.42	62.00	78.65	79.00
Our Approach	**64.04**	**64.00**	83.89	83.00

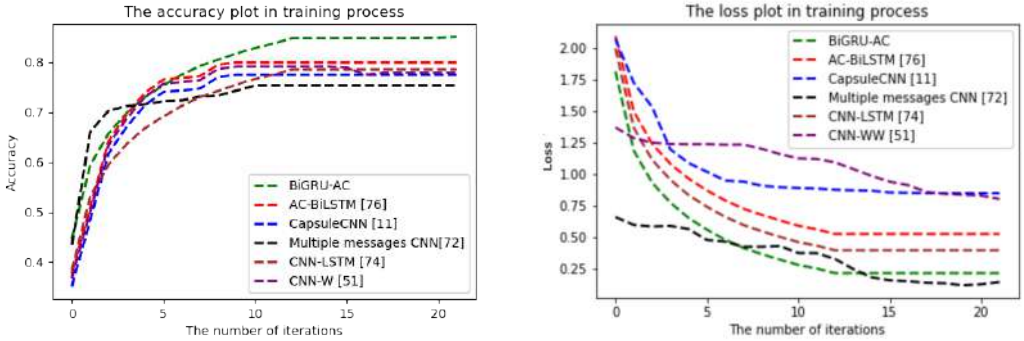
Table 15: Results for 10 authors

Methods	Blog10		Twitter10	
	acc	f_β	acc	f_β
Imposters [Koppel *et al.* 2011]		35.40		71.40
LDAH-S [Seroussi *et al.* 2011]		52.00		90.00
CNN [Ruder *et al.* 2016b]		61.		97.50
CNN-word-char [Ruder *et al.* 2016b]	60.10	60.00	95.50	95.00
Syntax-CNN [Zhang *et al.* 2018c]		64.10		
BertAA [Fabien *et al.* 2020a]		65.40		
ARC [Wen and Li 2018]	70.90	70.00	71.30	71.00
CNN-LSTM [Wang *et al.* 2016a]	66.02	64.00	85.70	86.00
CNN-LSTM [Wang *et al.* 2016a]	87.10	87.00	72.29	72.00
IWV [Rezaeinia *et al.* 2019]	73.50	73.00	73.50	73.00
AC-BLSTM [Liu and Guo 2019]	84.50	84.00	82.89	83.00
Our Approach	**85.79**	**85.00**	85.00	86.00

our ABC model intrinsically leverages the consequential effects of uniform resource locators, hashtags, and mentions of names in the posts to achieve high accuracy and F1 scores. Another point in Figure 45 (b) is the improvement of the accuracy of the proposed model for the correct classification of genuine authors for an input text compared to the loss for Twitter50 datasets. This may be because capturing the semantic relatedness among the features as the actual writing style is more challenging than identifying the author's vocabulary choice (e.g., domain-indicative lexical items)–more prevalent in the texts.

6.5.3.2 *Receiver operating characteristics (ROC) curve*

To better interpret the comprehensive evaluation performance of the proposed model, the receiver operating characteristics (ROC) curves of the ABC model were plotted as indicated in Figure 46 for IMDb62, Blog and Twitter dataset with 10 authors. The ROC helps to visualize

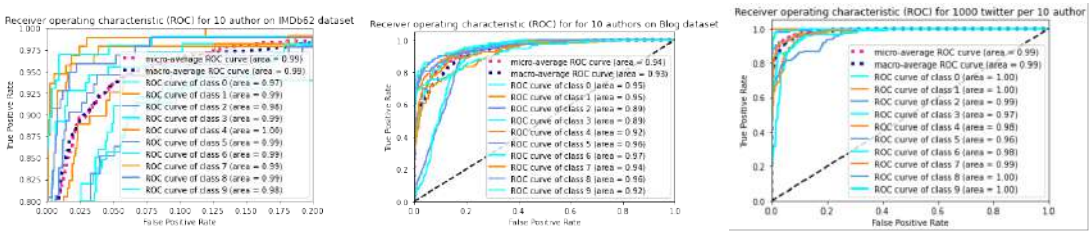


(a) Accuracy (b) Loss

Figure 45: Performance accuracy and loss curves over epoch on the Twitter50 datasets.

the performance of the proposed model using various thresholds [Fawcett 2006]. The area under the curve (AUC) ROC was employed as a classifier, based on the likelihood that the proposed model could rank a randomly positive instance higher than a randomly negative one on all the datasets when the class has ten authors.

A notable point in the results is that the performance improvement of the proposed model was consistent for two of the three datasets with 10 authors but lower for the blog dataset. This may be attributable to the average length of the sentences in these datasets. Figure 11 indicates the blog and Twitter datasets with fewer sentences than those in the IMDb datasets. As depicted in Figure 39, the proposed ABC model uses two RNNs, such as BLSTM and BGRU layers, on the embedding matrices to extract long dependencies in the input from online snippets. These indicate that the long dependencies are more evident in IMDb and Twitter datasets with a micro-average classification of over 90%, with the blog dataset near 94%; this demonstrates the better performance of the proposed ABC model on the AUC-ROC curve for IMDb and Twitter datasets compared with the blog dataset.



(a) IMDb62 (b) Blog (c) Twitter

Figure 46: ROC curve for IMDB62, Blog and Twitter datasets with 10 authors.

6.5.3.3 *Confusion matrix*

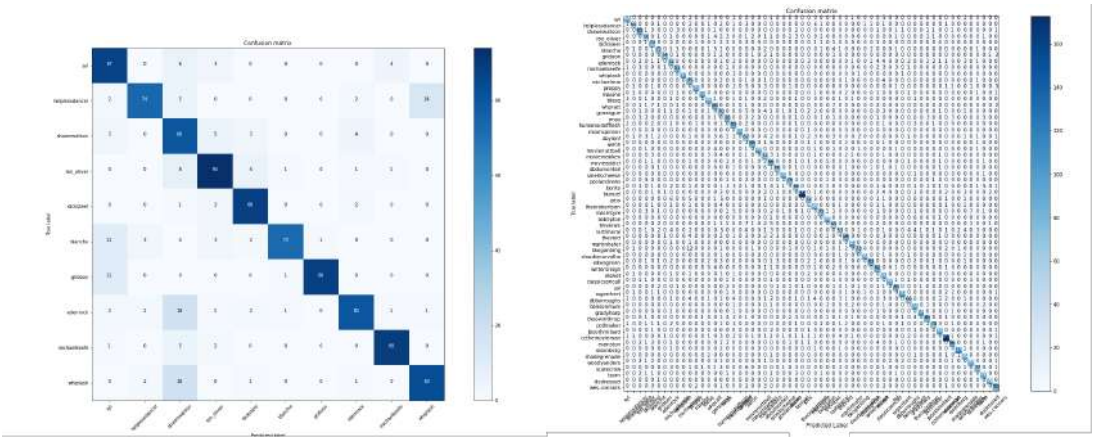
This same behavior is observed using a confusion matrix (CM) [Marom *et al.* 2010] to compute the error matrix of the proposed model at each iteration as provided in Figure 47, Figure 48 and Figure 49. A CM demonstrates the distribution of predictions over all classes in one concise observation. The percentage of calculating incorrect prediction for each class corresponding to each row is obtained by row-wise normalizing of the CM. The CM or cells on the diagonal demonstrated the normalized confusion matrix, indicating the recall for the corresponding class.

While the normalized matrix is useful for finding the percentage of the true and the false prediction, the CM with the actual number of counting the true and the false prediction

provides information on the size of each class in imbalanced data. Equation 47 show the calculation of CM used in the proposed ABC model.

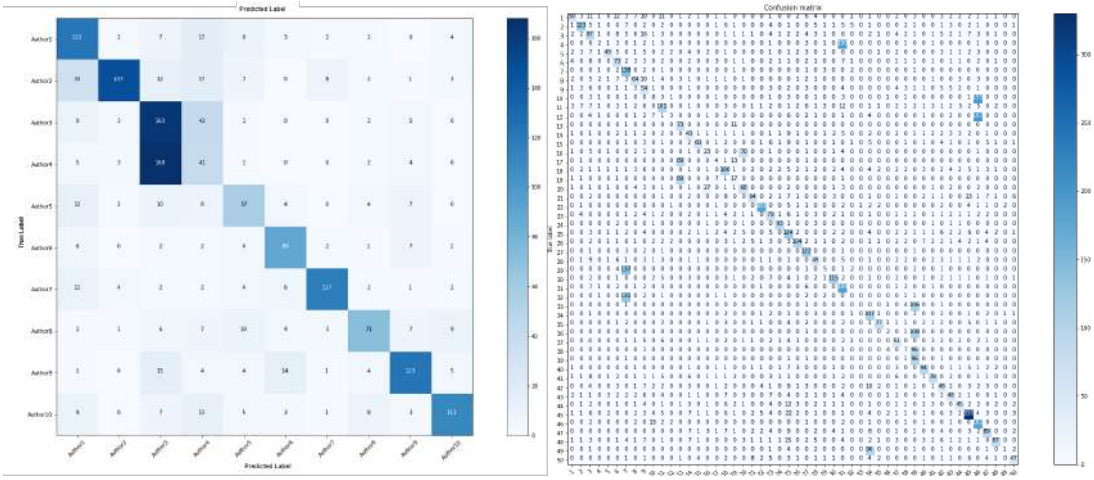
$$\mathrm{CM}(r,c) = \sum_{i=1}^{m} (\mathrm{I}(y_i = r)\mathrm{I}(h(x_i) = c)) \quad \forall_T, c \in \{0, \cdots, q-1\} \tag{47}$$

where CM is the confusion matrix, r and c are the row and column of the CM, m is the number of instances in the testing datasets, $I(\odot)$ represent the indicator function, and x_i is the i-th input to the proposed model $h(\odot)$, y_{i} represents the true class assigned to input x_i, and q is the number of class (or number of author).



(a) 10 authors (b) 62 authors

Figure 47: Confusion matrix for IMDb62 datasets with 10 & 62 authors.

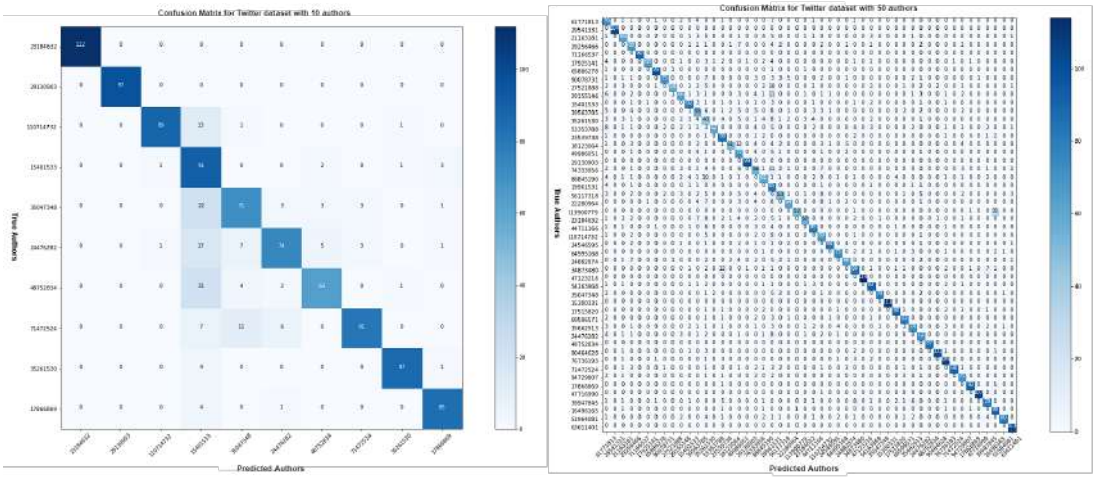


(a) 10 authors (b) 50 authors

Figure 48: Confusion matrix for Blog datasets with 10 & 50 authors.

6.5.4 *Result Visualization*

The visualization of results is useful for demonstrating how the feature vectors obtained from the proposed method can support specialists in understanding an author's writing style during the pretrial forensic investigation. The visualization analysis was performed using the local interpretable model-agnostic explanations (LIME) framework for interpretability [Lundberg and Lee 2017]. LIME explains how the proposed ABC method made predictions by identifying



(a) 10 authors (b) 50 authors

Figure 49: Confusion matrix for Twitter datasets with 10 & 50 authors.

the important input features.

Figure 50, Figure 51, and Figure 52 display the predictions from randomly selected sampled documents in the IMDb62, Blog50, and Twitter50 datasets. When compared to other authors, the ABC model learned the writing style of "Samuel Perry " with a 41% probability, as shown in Figure 50. It also contributed features like Hewlett-Packard, computers, prices, and technology to the online text. In Figure 52, the most important writing style to attribute the text sample to MartainWork' is the web, Internet, and company with 21%, while the other authors share 59% of the example documents.

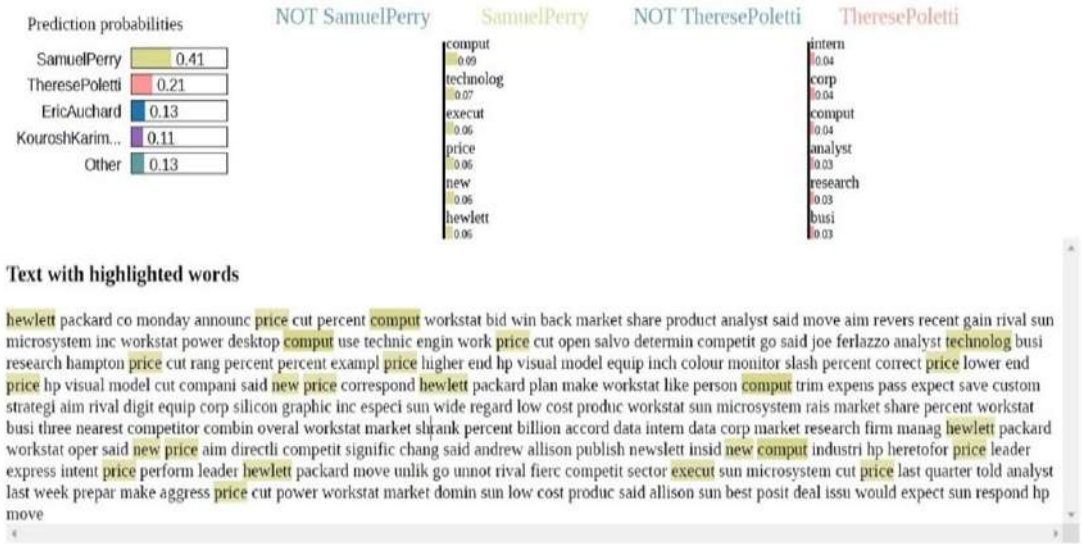


Figure 50: Explanation of individual prediction of the proposed ABC model on the IMDb50 dataset using LIME.

It was observed for the Twitter50 dataset in Figure 17 that the RDNN method assigned the same prediction probability of 32% to the authors JacaNews and Max Boot. The method put the same weight on the function words "American" and "Britain", representing two classes of authors, not JacaNews and JacaNews. The correct decision of the method demonstrates that

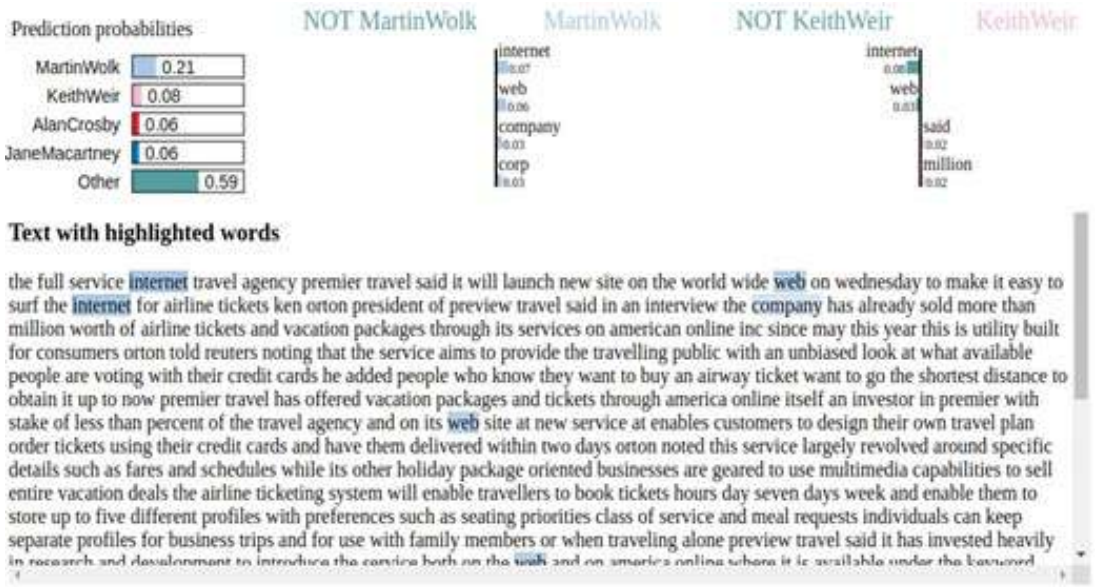


Figure 51: Explanation of individual prediction of the proposed ABC model on the Blog 50 dataset using LIME.

features such as "America" and "wit" are from JacaNews as an author, and words, such as "Britain" are more significant to Max Boot as the author with a probability value.

Figure 52: Explanation of individual prediction of the proposed ABC model on the Twitter50 dataset using LIME.

6.5.5 *Chapter Summary*

Post-authorship attribution is still an open research problem because of the intrinsic ambiguities in social media texts. As the initial weights of the embedding layer, the ABC model used pre-trained word2vec and GloVe word embedding vectors that were available to the public. On top of the embedding layer, two bidirectional LSTM and GRU networks extract past and future contexts as semantic representations from a post to identify the authorship of questionable texts. These semantic representations were passed to a convolutional layer comprising various kernel sizes to generalize various future maps and reduce the dimensionality of the feature space. An attention mechanism was applied to the outputs of the BLSTM and BGRU layers to pay more attention and learn the writing styles of the author across the datasets.

This enabled the proposed method to efficiently extract semantic representations, and more informative to identify the authorship of questionable texts; however, the rationale behind CNN was to boost the ABC model to learn local lexical and syntactic features and those long

dependencies established in the text by BLSTM and BGRU layers. Maximum and average pooling layers were stacked independently on the output of the CNN layers to ensure more robust resulting feature maps to learn positional changes in the author's writing style.

Finally, an FCL with a softmax function in the classification layer transformed the feature vector into an elevated level to predict the author of a text or structured document tag (SDT) based on the writing style representations. Experiments were conducted on two publicly available datasets and a new Twitter dataset to evaluate the model's performance. Experimental results on these datasets confirm that ABC achieves SOTA results on long and short-digit text.

The comparison of the results indicates that the improvements in the long text were less significant than in similar cases for online text snippets. The main reason may be that the feature layer in the ABC model was similar to the RNN-based network–designed to understand long sequential dependencies in the text better. The ABC was developed for the English language AA tasks; however, it could be easily extended to other languages to identify the text's authorship. Further analysis is desirable by introducing a stack generalization algorithm to improve the performance of the proposed method.

Part VII

CONCLUSIONS & FUTURE WORK

Chapter 7 discusses the main conclusion drawn from this study and proposes possible future work directions.

7 CONCLUSION AND FUTURE WORK

This chapter is the conclusion of the thesis that looked at how regularized deep neural networks can be used to use the writing styles of tweets, blogs, news articles, and documents to determine who wrote different kinds of ambiguous text snippets. The chapter provides a more detailed summary of the findings and contributions made throughout the thesis. It also demonstrates recommendations for future research in this area.

7.1 SUMMARY

Identification of authors and post-authorship attribution (PAA) tasks include looking at long documents, texts from terrorists that are questionable or suspicious, and online text snippets like tweets, blog posts, emails, forum posts, or chat logs to determine the genuine author of the text. We have designed and implemented an authorship identification system to help a criminal investigator, a forensic language specialist, and a qualified prosecutor gather empirical evidence by systematically analyzing large repositories of questionable (or suspicious) documents that came from many attempts to verify online information. The proposed methods employ deep regularized neural networks to learn the hidden writing style of a specific author from different intrinsic ambiguities in text snippets for improving the performance of post-authorship attribution and advanced authorship-based tasks.

The main goal of the thesis, as mentioned in Chapter 1, was to present more precise guidance for post-authorship attribution and possible ways of advancing research on authorship-based tasks by employing character-level, sub-word, and averaging word embedding to jointly learn a set of stylometric features such as lexical, syntactic, and semantics over a deep regularized neural network from a given post to identify the authorship of questionable texts. In Chapter 1, we believe that we have been able to tackle the following fact about the problem: many customised approaches focus on different sub-problems and scenarios of AA tasks. Therefore, we believe that we have been able to achieve this goal by tackling the following sub-problems:

- It is incredibly difficult to determine which stylometric features are most useful for extracting stylistic features related to writing style so that research in the PAA and AA systems can be improved and moved forward;
- The small size of the corpus could affect how well the style is represented. For instance, an inadequate sample of the dataset makes over-fitting worse, and a discrete or continuous representation may not be able to capture stylistic attributes in a small text dataset;
- Evaluation of the effectiveness of the proposed regularized deep neural network model from character-level, sub-word, and average word embedding to capture stylistic information across various types of datasets;
- The capacity to understand influential stylistic characteristics through a local explanation can be intuitively interpreted as a useful tool for forensic investigators during a pre-trial investigation.

Chapter 2 provides a review of previous research work on post-authorship attribution, either closed-set or open-set, which can be disintegrated into numerous authorship analyses, one for each author included in the potential candidate's list.

Chapter 3 looks at the features of human language using exploratory statistical analysis to see how language varies or is spread out across different types of datasets. In this chapter, we

tried to find words and phrases that make clear distinctions between groups and empathically point out events in the text so that we could learn more about the authors' writing styles. We then visualized and displayed across-associated words and phrases that are semantically similar to the author's deep, insightful writing style using open-source tools.

Chapter 4 investigates how well regularised deep neural networks (RDNN) model the inherent subjectivity in natural language. We did this to improve the performance of post-authorship attribution and move authorship-related task research forward. In the first part of our contribution, we show a convolutional neural network, a bidirectional long short-term memory encoder, and a distributed highway network that can pull out stylometric representations from different types of datasets to figure out who wrote documents that are in question or are clear.

Secondly, we proved that the lexical stylometric features fed into the bidirectional encoder to extract a syntactic feature vector representation were then supplied as input to the highly distributed high-speed networks to reduce the network generalisation error and gather feature vectors across a variety of datasets. We were able to correctly identify and capture the unique writing style of an author from an online text snippet and document by using regularisation. This has been a difficult task in computational linguistics and in previous research [Peng *et al.* 2003a; Hu *et al.* 2020b; Sari *et al.* 2017]. Thus, the regularised feature vector was subsequently implemented in the bidirectional decoder to acquire knowledge of the writing style of an author. This knowledge can then be used for various applications, such as authorship attribution or plagiarism detection.

Our ultimate contribution in this section is the introduction of a feature-classification layer, which consists of a fully connected network and a SoftMax function to make the prediction based on higher topical diversity and/or distinct idiosyncrasies in writing style to predict the actual author of the text or document. The proposed regularised neural network model was tested against SOTA methods using four standard experimental datasets to show that it worked well. Experimental results have demonstrated the effectiveness of the method when compared to the existing state-of-the-art methods on three datasets while producing comparable results on one dataset. Furthermore, we present the introduction of an interactive system to visualise the influence of the extracted writing style across the different datasets, which could be useful in the process of an evidence-based forensic investigation. This system allows users to explore and analyse the writing style patterns in various datasets, enhancing the efficiency and accuracy of forensic investigations.

Chapter 5 proposes combining a bidirectional long-term short-term memory (BLSTM) with a 2-dimensional convolutional neural network (CNN) to minimize writing styles that do not share the same vector space. Instead, we used subtoken embedding to deal with the ambiguity of words that were hidden in text documents while keeping the context of words or phrases that were used in different ways. To do this, the subordinate parts of words from different sequences of text in the corpus were taken. Then, use a sub-word embedding as a byte-pair encoding (BPE) to take advantage of overall compositions to reduce the size of the vocabulary needed to train the BLSTM-2DCNN model. Add Gaussian noise to make it easier to model the writer of a given text, which could give deep insights into the author's writing style. Experiments were done to compare the proposed BLSTM-2DCNN to many SOTA methods across the public corpus of datasets. The results of the experiments show that accuracy has improved on CCAT50 and Twitter, and the other datasets show similar results.

Chapter 6 presents an attention-based bidirectional long short-term memory and bidirectional gated recurrent unit with a convolutional neural network (ABC) model to circumvent post-authorship attribution challenges. It is based on a combination of bidirectional long short-term and gated recurrent units that extract a lexical stylometric from a distributed

representation of words that are fed into an attention-based mechanism to emphasize syntactic feature vector representation categories. The feature vector is input into a convolution and pooling mechanism, reducing the feature dimensionality while learning the position-invariant of the writing style. A Softmax function predicted the feature classification. The ABC method was evaluated against trendsetting methods using four benchmark experimental datasets. The results showed that it outperformed the methods on two datasets while producing comparable effects on the other two.

In this thesis, we evaluated the proposed regularized deep neural network for post-authorship attribution and style-related tasks to advance research on authorship-based tasks on various datasets. We demonstrated how each of the proposed models works in various situations, as well as how to extract a stylometric representation vector to achieve a certain level of reliability in writing style understanding and improve performance. The outcomes of this thesis demonstrate that the influential features of the writing style that contributed to the authors can be visualized and used as language-based evidence during the pre-trial investigation. Additionally, we have provided suggestions on the proposed models and features that produce optimal performance given a certain type of dataset.

7.2 FUTURE WORK

A diversity of post-authorship attribution and other style-related tasks are explored in this thesis. Similarly, the proposed methods can be extended into different domains, as outlined as follows:

- **Datasets for generic authorship attribution are required**: One of the critical aspects of future authorship attribution work is the creation of a generic set of data that will be measured and evaluated by every developed application or technique. Thus, the lack of standard datasets for a different type of authorship algorithm hampered the pursuit of more sophisticated tools using deep neural networks attributed to the inaccessibility of datasets with enormous numbers of authors in the training examples. Notwithstanding, building an authorship attribution dataset is a very challenging phenomenon. For instance, [Juola and Vescovi 2010; Rudman 2012] highlighted the valuable contribution of building a dataset from an unadulterated and genuine piece of the author's writings. Therefore, we presume standard datasets will strengthen the continuation of systematic work in the authorship attribution field.

- **Training data diversity**: In recent years, there is a diversity in the way humans communicate with a social group. However, the majority of authorship attribution systems employed English datasets to evaluate the approach, except very few [Peng *et al.* 2003a; Mikros and Perifanos 2013]. There should be heterogeneity sampling in developing experimental datasets to help develop the most appropriate useful algorithm that preserves writing style varies across all domains of languages. Although a deep neural network needs a large dataset and is time-consuming to train, sometimes the best way to make a meaningful decision from the data is to extract much smaller subsets and develop an unsupervised machine learning system to select concise stylometric representation from all the range of cases the system will have to confront. Another potential path is to construct an adaptive approach employed for any language.

- **Exploring transfer learning**: The inspiration comes from us – humans, ourselves – wherein, we have an inherent ability not to learn everything from scratch. The method is a fascinating phenomenon where forensic authorship analysis system seeks to exploit prior knowledge from one domain and task to some other areas and tasks. The idea is to use generic embedding (e.g., pre-trained models) [Lu *et al.* 2015; Ruder 2019] to investigate tasks such as authorship obfuscation, imitation, and translation [Brennan *et al.* 2012]. The method read input text by encoding them into a fixed-dimensional embedding vector trained on a variety of natural language understanding tasks that model the

interpretation of the sequences in the text rather than just memories of the individual words. An alternative is to use neural style transfer on massive amount of dataset to reconstruct obfuscated writing style from original passages [Shen *et al.* 2017; Jing *et al.* 2019].

BIBLIOGRAPHY

[Abadi *et al.* 2013] Daniel Abadi, Peter Boncz, Stavros Harizopoulos, Stratos Idreos, Samuel Madden, et al. The design and implementation of modern column-oriented database systems. *Foundations and Trends® in Databases*, 5(3):197–280, 2013.

[Abbasi and Chen 2005] Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[Abbasi and Chen 2008] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7, 2008.

[Abbasi *et al.* 2008] Ahmed Abbasi, Hsinchun Chen, and Jay F Nunamaker. Stylometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, 25(1):49–78, 2008.

[Abbass *et al.* 2020] Zaheer Abbass, Zain Ali, Mubashir Ali, Bilal Akbar, and Ahsan Saleem. A framework to predict social crime through twitter tweets by using machine learning. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 363–368, 2020.

[Abdi and Williams 2010] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[Abelló *et al.* 2008] Alberto Abelló, M Elena Rodríguez, Toni Urpí, Xavier Burgués, M José Casany, Carme Martín, and Carme Quer. LEARN-SQL: Automatic assessment of SQL based on IMS QTI specification. In *Eighth IEEE International Conference on Advanced Learning Technologies*, pages 592–593. IEEE, 2008.

[Abiteboul *et al.* 2005] Serge Abiteboul, Rakesh Agrawal, Phil Bernstein, Mike Carey, Stefano Ceri, Bruce Croft, David DeWitt, Mike Franklin, Hector Garcia Molina, Dieter Gawlick, et al. The Lowell database research self-assessment. *Communications of the ACM*, 48(5):111–118, 2005.

[Aborisade and Anwar 2018] Opeyemi Aborisade and Mohd Anwar. Classification for authorship of tweets by comparing logistic regression and naive bayes classifiers. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 269–276, July 2018.

[Acquisti and Gross 2006] Alessandro Acquisti and Ralph Gross. Imagined communities: Awareness, information sharing, and privacy on the facebook. In George Danezis and Philippe Golle, editors, *Privacy Enhancing Technologies*, pages 36–58, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[Adams 2015] Nancy E Adams. Bloomâ€™s taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, 103(3):152, 2015.

[Adams 2017] Marilyn Jager Adams. Failures to comprehend and levels of processing. *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence and education*, 11:11, 2017.

[Adegbola 2021] Tunde Adegbola. Linguistic diversity in the age of big data. *Language Technologies and Language Diversity Tecnologies de la llengua i diversitat lingüística*, page 63, 2021.

[af Segerstad 2002] Ylva Hård af Segerstad. *Use and adaptation of written language to the conditions of computer-mediated communication*. Department of Linguistics, Univ., 2002.

[Affolter *et al.* 2019] Katrin Affolter, Kurt Stockinger, and Abraham Bernstein. A comparative survey of recent natural language interfaces for databases. *arXiv preprint arXiv:1906.08990*, 2019.

[Afroz *et al.* 2012] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*, pages 461–475, May 2012.

[Afroz *et al.* 2014] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. Doppelgänger finder: Taking stylometry to the underground. In *2014 IEEE Symposium on Security and Privacy*, pages 212–226. IEEE, 2014.

[Agarwal and Kumar 2016] Deepak Kumar Agarwal and Rahul Kumar. Spam filtering using SVM with different kernel functions. *International Journal of Computer Applications*, 136(5):0975–8887, 2016.

[Agarwal *et al.* 2012] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S Yu. Modeling blogger influence in a community. *Social Network Analysis and Mining*, 2(2):139–162, 2012.

[Aggarwal and Zhai 2012] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer, 2012.

[Aggarwal *et al.* 2012] Anupama Aggarwal, Ashwin Rajadesingan, and Ponnurangam Kumaraguru. Phishari: Automatic realtime phishing detection on twitter. In *2012 eCrime Researchers Summit*, pages 1–12, 2012.

[Agrawal *et al.* 2018] Amritanshu Agrawal, Wei Fu, and Tim Menzies. What is wrong with topic modeling? and how to fix it using search-based software engineering. *Information and Software Technology*, 98:74–88, 2018.

[Ahadi *et al.* 2015] Alireza Ahadi, Julia Prior, Vahid Behbood, and Raymond Lister. A quantitative study of the relative difficulty for novices of writing seven different types of SQL queries. In *Proceedings of the ACM Conference on Innovation and Technology in Computer Science Education*, pages 201–206. ACM, 2015.

[Ahadi *et al.* 2016] Alireza Ahadi, Vahid Behbood, Arto Vihavainen, Julia Prior, and Raymond Lister. Students' syntactic mistakes in writing seven different types of SQL queries and its application to predicting students' success. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pages 401–406. ACM, 2016.

[Aho *et al.* 1986] Alfred V Aho, Ravi Sethi, and Jeffrey D Ullman. Compilers, principles, techniques. *Addison Wesley*, 7(8):9, 1986.

[Aken and Michalisin 2007] Andrew Aken and Michael D Michalisin. The impact of the skills gap on the recruitment of MIS graduates. In *Proceedings of the 2007 ACM SIGMIS CPR conference on Computer personnel research: The global information technology workforce*, pages 105–111. ACM, 2007.

[Akiva and Koppel 2012] Navot Akiva and Moshe Koppel. Identifying distinct components of a multi-author document. In *2012 European Intelligence and Security Informatics Conference*, pages 205–209, Aug 2012.

[Akiva and Koppel 2013] Navot Akiva and Moshe Koppel. A generic unsupervised method for decomposing multi-author documents. *Journal of the American Society for Information Science and Technology*, 64(11):2256–2264, 2013.

[Al-Ahmad *et al.* 2021] Bilal Al-Ahmad, Alaâ€™M Al-Zoubi, Ruba Abu Khurma, and Ibrahim Aljarah. An evolutionary fake news detection method for covid-19 pandemic information. *Symmetry*, 13(6):1091, 2021.

[Al-Emran and Malik 2016] Mostafa Al-Emran and Sohail Iqbal Malik. The impact of google apps at work: Higher educational perspective. *Int. J. Interact. Mob. Technol.*, 10(4):85–88, 2016.

[Al-Garadi *et al.* 2016] Mohammed Ali Al-Garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433–443, 2016.

[Al Omran and Treude 2017] Fouad Nasser A Al Omran and Christoph Treude. Choosing an NLP library for analyzing software documentation: a systematic literature review and a series of experiments. In *Proceedings of the 14th International Conference on Mining Software Repositories*, pages 187–197. IEEE Press, 2017.

[Al-Radaideh and Al-Abrat 2019] Qasem A Al-Radaideh and Mohammed A Al-Abrat. An Arabic text categorization approach using term weighting and multiple reducts. *Soft Computing*, 23(14):5849–5863, 2019.

[Al-Shuaily and Renaud 2010] Huda Al-Shuaily and Karen Renaud. SQL patterns-a new approach for teaching SQL. In *8th HEA Workshop on Teaching, Learning and Assessment of Databases, Abertay-Dundee*, pages 29–40, 2010.

[Al Shuaily and Renaud 2016a] Huda Al Shuaily and Karen Renaud. A framework for SQL learning: Linking learning taxonomy, cognitive model and cross cutting factors. *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 10(9):3091–3097, 2016.

[Al Shuaily and Renaud 2016b] Huda Al Shuaily and Karen Renaud. A model of SQL learning. 2016.

[Al-Shuaily 2013] Huda Al-Shuaily. *SQL pattern design, development & evaluation of its efficacy*. PhD thesis, University of Glasgow, 2013.

[Al Zamal *et al.* 2012] Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 387–390, 2012.

[Alabid and Katheeth 2021] Noralhuda N Alabid and Zainab Dalaf Katheeth. Sentiment analysis of twitter posts related to the covid-19 vaccines. *Indonesian Journal of Electrical Engineering and Computer Science*, 24(3):1727–1734, 2021.

[Aldahdooh and Naser 2017] Rami Aldahdooh and Samy S. Abu Naser. Development and evaluation of the oracle intelligent tutoring system (oits). *European Academic Research*, 4(10), 2017.

[Alexander 1977] Christopher Alexander. *A pattern language: towns, buildings, construction*. Oxford university press, 1977.

[Aljumily 2015] Refat Aljumily. Hierarchical and non-hierarchical linear and non-linear clustering methods to shakespeare authorship question. *Social Sciences*, 4(3):758–799, 2015.

[Allenstein *et al.* 2008] Brett Allenstein, Andrew Yost, Paul Wagner, and Joline Morrison. A query simulation system to illustrate database query execution. *ACM Special Interest Group on Computer Science Education Bulletin*, 40(1):493–497, 2008.

[Allison and Guthrie 2008] Ben Allison and Louise Guthrie. Authorship attribution of e-mail: Comparing classifiers over a new corpus for evaluation. In *LREC*, 2008.

[Almishari *et al.* 2014] Mishari Almishari, Ekin Oguz, and Gene Tsudik. Fighting authorship linkability with crowdsourcing. In *Proceedings of the second ACM conference on Online social networks*, pages 69–82, 2014.

[Alonso-Fernandez *et al.* 2021] Fernando Alonso-Fernandez, Nicole Mariah Sharon Belvisi, Kevin Hernandez-Diaz, Naveed Muhammad, and Josef Bigun. Writer identification using microblogging texts for social media forensics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):405–426, 2021.

[Alpaydin 2020] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[Alsajjan and Dennis 2010] Bander Alsajjan and Charles Dennis. Internet banking acceptance model: Cross-market examination. *Journal of Business Research*, 63(9):957–963, 2010.

[AlSallal *et al.* 2019] Muna AlSallal, Rahat Iqbal, Vasile Palade, Saad Amin, and Victor Chang. An integrated approach for intrinsic plagiarism detection. *Future Generation Computer Systems*, 96:700–712, 2019.

[Alshammari and Nielsen 2018] Sultanah M Alshammari and Rodney D Nielsen. Less is more: with a 280-character limit, twitter provides a valuable source for detecting self-reported flu cases. In *Proceedings of the 2018 International Conference on Computing and Big Data*, pages 1–6, 2018.

[Alshamrani *et al.* 2020] Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Anho Lee, DaeHun Nyang, and David Mohaisen. An analysis of users engagement on twitter during the covid-19 pandemic: Topical trends and sentiments. In Sriram Chellappan, Kim-Kwang Raymond Choo, and NhatHai Phan, editors, *Computational Data and Social Networks*, pages 73–86, Cham, 2020. Springer International Publishing.

[Alsohybe *et al.* 2017] Nabeel T Alsohybe, Neama Abdulaziz Dahan, and Fadl Mutaher Ba-Alwi. Machine-translation history and evolution: Survey for Arabic-English translations. *arXiv preprint arXiv:1709.04685*, 2017.

[Altakrori *et al.* 2018] Malik H Altakrori, Farkhund Iqbal, Benjamin CM Fung, Steven HH Ding, and Abdallah Tubaishat. Arabic authorship attribution: An extensive study on twitter posts. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(1):1–51, 2018.

[Amodei *et al.* 2016] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[Andreeva *et al.* 2018] Ol'ga Vyacheslavovna Andreeva, Mirabbas Bakhtiyarovich Bagirov, Anastasiya Aleksandrovna Dankina, Tat'yana Olegovna Fedorova, and Mariya Mikhailovna Sheveleva. Intellectual analysis of data on the basis of Stanford coreNLP for PoS tagging of texts in the Russian language. *Sistemy i Sredstva Informatiki [Systems and Means of Informatics]*, 28(2):145–153, 2018.

[Andriole 2010] Stephen J Andriole. Business impact of web 2.0 technologies. *Communications of the ACM*, 53(12):67–79, 2010.

[Angeli *et al.* 2014] Gabor Angeli, Sonal Gupta, Melvin Johnson Premkumar, Christopher D Manning, Christopher Ré, Julie Tibshirani, Jean Y Wu, Sen Wu, and Ce Zhang. Stanfordâ€™s distantly supervised slot filling systems for kbp 2014. In *Proceedings of the Seventh Text Analysis Conference*, 2014.

[Anwar *et al.* 2018] Waheed Anwar, Imran Sarwar Bajwa, M Abbas Choudhary, and Shabana Ramzan. An empirical study on forensic analysis of urdu text using lda-based authorship attribution. *IEEE Access*, 7:3224–3234, 2018.

[Apoorva and Sangeetha 2021] K A Apoorva and S Sangeetha. Forensic analysis of e-mail for authorship attribution: Research perspective. In *Proceeding of First Doctoral Symposium on Natural Computing Research*, pages 281–292. Springer, 2021.

[Apostolato 2013] Ioana-Alexandra Apostolato. An overview of software applications for social network analysis. *International Review of Social Research*, 3(3), 2013.

[Arampatzi *et al.* 2018] Efstratia Arampatzi, Martijn J Burger, and Natallia Novik. Social network sites, individual social capital and happiness. *Journal of Happiness Studies*, 19:99–122, 2018.

[Arasu *et al.* 2016] Arvind Arasu, Brian Babcock, Shivnath Babu, John Cieslewicz, Mayur Datar, Keith Ito, Rajeev Motwani, Utkarsh Srivastava, and Jennifer Widom. Stream: The Stanford Data Stream Management System. In *Data Stream Management*, pages 317–336. Springer, 2016.

[Ardito *et al.* 2014] Carmelo Ardito, Maria Francesca Costabile, and Hans-Christian Jetter. Gestures that people can understand and use. *Journal of Visual Languages & Computing*, 25(5):572–576, 2014.

[Argamon and Levitan 2005] Shlomo Argamon and Shlomo Levitan. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, pages 4–7, 2005.

[Argamon *et al.* 2007] Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, 2007.

[Argamon *et al.* 2009] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[Argamon *et al.* 2010] Shlomo Argamon, Kevin Burns, and Shlomo Dubnov. *The structure of style: Algorithmic approaches to understanding manner and meaning*. Springer, 2010.

[Argamon-Engelson *et al.* 1998] Shlomo Argamon-Engelson, Moshe Koppel, and Galit Avneri. Style-based text categorization: What newspaper am i reading. In *Proc. of the AAAI Workshop on Text Categorization*, pages 1–4, 1998.

[Armbrust *et al.* 2015] Michael Armbrust, Reynold S Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K Bradley, Xiangrui Meng, Tomer Kaftan, Michael J Franklin, Ali Ghodsi, et al. Spark SQL: Relational data processing in spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1383–1394. ACM, 2015.

[Artetxe *et al.* 2017] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*, 2017.

[Arun *et al.* 2009] Rajkumar Arun, Ravi Saradha, V Suresh, M Murty, and C Madhavan. Stopwords and stylometry: a latent dirichlet allocation approach. In *NIPS workshop on Applications for Topic Models*, 2009.

[Atchariyachanvanich *et al.* 2017] Kanokwan Atchariyachanvanich, Srinual Nalintippayawong, and Tanasab Permpool. Development of a mySQL sandbox for processing SQL statements: Case of DML and DDL statements. In *2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–6. IEEE, 2017.

[Athenikos and Han 2010] Sofia J Athenikos and Hyoil Han. Biomedical question answering: A survey. *Computer methods and programs in biomedicine*, 99(1):1–24, 2010.

[Auria and Moro 2008] Laura Auria and Rouslan A Moro. Support vector machines (SVM) as a technique for solvency analysis. (881):1–18, 2008.

[Awad and Khanna 2015] Mariette Awad and Rahul Khanna. *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Apress, 2015.

[Aykent and Dozier 2020] Sarp Aykent and Gerry Dozier. Author identification of micro-messages via multi-channel convolutional neural networks. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 675–681. IEEE, 2020.

[Azarbonyad *et al.* 2015] Hosein Azarbonyad, Mostafa Dehghani, Maarten Marx, and Jaap Kamps. Time-aware authorship attribution for short text streams. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 727–730, 2015.

[Azer *et al.* 2013] Samy A Azer, Michelle Mclean, Hirotaka Onishi, Masami Tagawa, and Albert Scherpbier. Cracks in problem-based learning: What is your action plan? *Medical teacher*, 35(10):806–814, 2013.

[Baayen *et al.* 1996] Harald Baayen, Hans Van Halteren, and Fiona Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132, 1996.

[Baayen 2001] R Harald Baayen. *Word frequency distributions*, volume 18. Springer Science & Business Media, 2001.

[Babur *et al.* 2022] Önder Babur, Loek Cleophas, and Mark van den Brand. Samos-a framework for model analytics and management. *Science of Computer Programming*, 223:102877, 2022.

[Badea 2014] Marius Badea. Social media and organizational communication. *Procedia - Social and Behavioral Sciences*, 149:70–75, 2014. LUMEN 2014 - From Theory to Inquiry in Social Sciences, Iasi, Romania, 10-12 April 2014.

[Bagnall 2015] Douglas Bagnall. *Author Identification using Multi-headed Recurrent Neural Networks*, 2015.

[Bagul and Barve 2021] Dhiraj Vaibhav Bagul and Sunita Barve. A novel content-based recommendation approach based on lda topic modeling for literature recommendation. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 954–961, 2021.

[Bahdanau *et al.* 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Bahdanau *et al.* 2016] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE, 2016.

[Baik *et al.* 2019] Christopher Baik, HV Jagadish, and Yunyao Li. Bridging the semantic gap with SQL query logs in natural language interfaces to databases. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 374–385. IEEE, 2019.

[Bakardjieva and Gaden 2012] Maria Bakardjieva and Georgia Gaden. Web 2.0 technologies of the self. In *Cultural Technologies*, pages 165–181. Routledge, 2012.

[Bakshy *et al.* 2011] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74, 2011.

[Balakrishnan *et al.* 2022] Vimala Balakrishnan, Ng Wei Zhen, Soo Mun Chong, Gan Joo Han, and Tan Jiat Lee. Infodemic and fake news–a comprehensive overview of its global magnitude during the covid-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction*, page 103144, 2022.

[Baldridge 2005] Jason Baldridge. The openNLP project. *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*, page 1, 2005.

[Barber *et al.* 2018] Mark H Barber, Carsten Hagemann, and Christopher J Hockings. *Similar email spam detection*, 2018. US Patent App. 15/270,237.

[Barlas and Stamatatos 2020a] Georgios Barlas and Efstathios Stamatatos. Cross-domain authorship attribution using pre-trained language models. In Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, pages 255–266, Cham, 2020. Springer International Publishing.

[Barlas and Stamatatos 2020b] Georgios Barlas and Efstathios Stamatatos. Cross-domain authorship attribution using pre-trained language models. In Ilias Maglogiannis, Lazaros Iliadis, and Elias Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, pages 255–266, Cham, 2020. Springer International Publishing.

[Baron 2004] Naomi S Baron. Why email looks like speech: Proofreading, pedagogy and public face. In *New media language*, pages 99–108. Routledge, 2004.

[Baron 2016] Grzegorz Baron. Comparison of cross-validation and test sets approaches to evaluation of classifiers in authorship attribution domain. In Tadeusz Czachórski, Erol Gelenbe, Krzysztof Grochla, and Ricardo Lent, editors, *Computer and Information Sciences*, pages 81–89, Cham, 2016. Springer International Publishing.

[Baroni *et al.* 2014] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Donâ€™t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, 2014.

[Barrera and Pachitariu 2018] Jorge Barrera and George Pachitariu. Big data: What is it? and is my data big enough? 2018.

[Barron *et al.* 2015] Andrew B Barron, Eileen A Hebets, Thomas A Cleland, Courtney L Fitzpatrick, Mark E Hauber, and Jeffrey R Stevens. Embracing multiple definitions of learning. *Trends in neurosciences*, 38(7):405–407, 2015.

[Bartelds and de Vries 2019] Martijn Bartelds and Wietse de Vries. Improving cross-domain authorship attribution by combining lexical and syntactic features. In *CLEF (Working Notes)*, 2019.

[Basile *et al.* 2019] Angelo Basile, Albert Gatt, and Malvina Nissim. You write like you eat: Stylistic variation as a predictor of social stratification. *arXiv preprint arXiv:1907.07265*, 2019.

[Bast and Haussmann 2015] Hannah Bast and Elmar Haussmann. More accurate question answering on freebase. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1431–1440. ACM, 2015.

[Bastani *et al.* 2017] Osbert Bastani, Rahul Sharma, Alex Aiken, and Percy Liang. Synthesizing program input grammars. In *ACM Special Interest Group on Programming Languages*, volume 52, pages 95–110. ACM, 2017.

[Bateman and Zock 2003] John Bateman and Michael Zock. Natural language generation. In *The Oxford Handbook of Computational Linguistics 2nd edition*. 2003.

[Bates 1995] Madeleine Bates. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982, 1995.

[Batini *et al.* 1991] Carlo Batini, Tiziana Catarci, Maria Francesca Costabile, and Stefano Levialdi. Visual query systems: A taxonomy. In *VDB*, pages 153–168, 1991.

[Batra 2018] Rahul Batra. A history of SQL and relational databases. In *SQL Primer*, pages 183–187. Springer, 2018.

[Bau 2015] David Bau. Droplet, a blocks-based editor for text code. *Journal of Computing Sciences in Colleges*, 30(6):138–144, 2015.

[Baur *et al.* 2002] Michael Baur, Marc Benkert, Ulrik Brandes, Sabine Cornelsen, Marco Gaertler, Boris Köpf, Jürgen Lerner, and Dorothea Wagner. Visone software for visual social network analysis. In *Graph Drawing: 9th International Symposium, GD 2001 Vienna, Austria, September 23–26, 2001 Revised Papers 9*, pages 463–464. Springer, 2002.

[Baykasoğlu *et al.* 2018] Adil Baykasoğlu, Burcu K Özbel, Nurhan Dudaklı, Kemal Subulan, and Mümin Emre Şenol. Process mining based approach to performance evaluation in computer-aided examinations. *Computer Applications in Engineering Education*, 26(5):1841–1861, 2018.

[Beaulieu 2009] Alan Beaulieu. *Learning SQL: Master SQL Fundamentals*. " O'Reilly Media, Inc.", 2009.

[Becker and Stamp 2005] Jennifer AH Becker and Glen H Stamp. Impression management in chat rooms: A grounded theory model. *Communication studies*, 56(3):243–260, 2005.

[Behari *et al.* 2016] Suren Behari, Aileen Cater-Steel, and Jeffrey Soar. Data science and big data analytics in financial services: A case study. *Handbook of Research on Driving Competitive Advantage through Sustainable, Lean, and Disruptive Innovation*, page 396, 2016.

[Belanus 2004] Betty J Belanus. Cyberpl@ y: Communicating online. *Journal of American Folklore*, 117(463):119–120, 2004.

[Belinkov *et al.* 2015] Yonatan Belinkov, Mitra Mohtarami, Scott Cyphers, and James Glass. Vectorslu: A continuous word vector approach to answer selection in community question answering systems. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 282–287, 2015.

[Belvisi *et al.* 2020] Nicole Mariah Sharon Belvisi, Naveed Muhammad, and Fernando Alonso-Fernandez. Forensic authorship analysis of microblogging texts using n-grams and stylometric features. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020.

[Ben-Hur and Weston 2010] Asa Ben-Hur and Jason Weston. A userâ€™s guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer, 2010.

[Bengio and Senécal 2008] Yoshua Bengio and Jean-Sébastien Senécal. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Transactions on Neural Networks*, 19(4):713–722, 2008.

[Bengio *et al.* 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[Benkhelifa *et al.* 2014] Elhadj Benkhelifa, Elliott Rowe, Robert Kinmond, Oluwasegun A Adedugbe, and Thomas Welsh. Exploiting social networks for the prediction of social and civil unrest: A cloud based framework. In *2014 international conference on future internet of things and cloud*, pages 565–572. IEEE, 2014.

[Benoit *et al.* 2019] Kenneth Benoit, Kevin Munger, and Arthur Spirling. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, 63(2):491–508, 2019.

[Bentley 1999] Jon Bentley. *Programming Pearls*. Addison–Wesley, Boston, MA, USA, 2nd edition, 1999.

[Berant and Liang 2014] Jonathan Berant and Percy Liang. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, 2014.

[Bergin *et al.* 2012] Joseph Bergin, Jutta Eckstein, M Manns, H Sharp, K Marquardt, Jane Chandler, M Sipos, M Volter, and E Wallingford. *Pedagogical patterns: advice for educators*. CreateSpace, 2012.

[Berninger *et al.* 2006] Virginia W Berninger, Robert D Abbott, Janine Jones, Beverly J Wolf, Laura Gould, Marci Anderson-Youngstrom, Shirley Shimada, and Kenn Apel. Early development of language by hand: Composing, reading, listening, and speaking connections; three letter-writing modes; and fast mapping in spelling. *Developmental neuropsychology*, 29(1):61–92, 2006.

[Berque *et al.* 2003] Dave Berque, Terri Bonebright, Seth Kinnett, Nathan Nichols, and Adam Peters. A case study in the design of software that uses auditory cues to help low vision students view notes on a blackboard. Georgia Institute of Technology, 2003.

[Berthold *et al.* 2009] Kirsten Berthold, Tessa HS Eysink, and Alexander Renkl. Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science*, 37(4):345–363, 2009.

[Bevendorff *et al.* 2019] Janek Bevendorff, Matthias Hagen, Benno Stein, and Martin Potthast. Bias analysis and mitigation in the evaluation of authorship verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6301–6306, 2019.

[Bevendorff *et al.* 2021] Janek Bevendorff, Berta Chulvi, Gretel Liz De La Peña Sarracén, Mike Kestemont, Enrique Manjavacas, Ilia Markov, Maximilian Mayerl, Martin Potthast, Francisco Rangel, Paolo Rosso, et al. Overview of pan 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 419–431. Springer, 2021.

[Bhargava *et al.* 2013] Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. Stylometric analysis for authorship attribution on twitter. In *Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, December 16-18, 2013, Proceedings 2*, pages 37–47. Springer, 2013.

[Bhuiyan *et al.* 2020] Mohd Salman Hossain Bhuiyan, Nabil Al Raian, Shahad Iqbal Leon, and Musharrat Khan. Study of influence of dimension reduction of high dimensional datasets in classification problem. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 147–151. IEEE, 2020.

[Bian *et al.* 2014] Jiang Bian, Bin Gao, and Tie-Yan Liu. Knowledge-powered deep learning for word embedding. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 132–148. Springer, 2014.

[Biau 2012] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

[Bider and Rogers 2016] Ilia Bider and David Rogers. YASQLT–Yet Another SQL Tutor. In *International Conference on Conceptual Modeling*, pages 197–206. Springer, 2016.

[Bilgin 2019] Metin Bilgin. A study on named entity recognition with openNLP at English texts. *Journal of Applied Intelligent System*, 4(1):1–8, 2019.

[Bin Naeem and Kamel Boulos 2021] Salman Bin Naeem and Maged N Kamel Boulos. Covid-19 misinformation online and health literacy: a brief overview. *International journal of environmental research and public health*, 18(15):8091, 2021.

[Bird *et al.* 2008] Steven Bird, Ewan Klein, Edward Loper, and Jason Baldridge. Multidisciplinary instruction with the natural language toolkit. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics*, pages 62–70. Association for Computational Linguistics, 2008.

[Bird *et al.* 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[Bird 2002] Steven Bird. Computational phonology. *arXiv preprint cs/0204023*, 2002.

[Bird 2006] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

[Blei *et al.* 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[Bobicev *et al.* 2013] Victoria Bobicev, Marina Sokolova, Khaled El Emam, Yasser Jafer, Brian Dewar, Elizabeth Jonker, Stan Matwin, et al. Can anonymous posters on medical forums be reidentified? *Journal of medical Internet research*, 15(10):e2514, 2013.

[Bobrow *et al.* 1977] Daniel G Bobrow, Ronald M Kaplan, Martin Kay, Donald A Norman, Henry Thompson, and Terry Winograd. Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173, 1977.

[Bocklisch *et al.* 2017] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

[Bogard 2004] Travis A Bogard. Instant messaging via telephone interfaces. *Google Patents*, 2004.

[Bogdanova and Lazaridou 2014] Dasha Bogdanova and Angeliki Lazaridou. Cross-language authorship attribution. In *LREC*, pages 2015–2020, 2014.

[Bojanowski *et al.* 2017] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[Bondielli *et al.* 2018] Alessandro Bondielli, Lucia C Passaro, and Alessandro Lenci. CoreNLP-it: A UD pipeline for italian based on Stanford coreNLP. In *CLiC-it*, 2018.

[Bonham-Carter 2014] Graeme F Bonham-Carter. *Geographic information systems for geoscientists: modelling with GIS*, volume 13. Elsevier, 2014.

[Boot *et al.* 2019] Arnout B Boot, Erik Tjong Kim Sang, Katinka Dijkstra, and Rolf A Zwaan. How character limit affects language usage in tweets. *Palgrave Communications*, 5(1), 2019.

[Borchers and Thomas 2001] Jan O Borchers and John C Thomas. Patterns: what's in it for HCI? In *CHI'01 extended abstracts on Human factors in computing systems*, pages 225–226. ACM, 2001.

[Borchers 2000] Jan O Borchers. Interaction design patterns: twelve theses. In *Workshop, The Hague*, volume 2, page 3. Citeseer, 2000.

[Bornat 1987] Richard Bornat. *Programming from First Principles*. Prentice Hall International (UK) Ltd., Hertfordshire, UK, UK, 1987.

[Borodin *et al.* 2016] Andrey Borodin, Yuri Kiselev, Sergey Mirvoda, and Sergey Porshnev. Development of data aggregation capabilities in domain-specific query language for metallurgy. In *2016 Dynamics of Systems, Mechanisms and Machines (Dynamics)*, pages 1–6. IEEE, 2016.

[Borrego 2007] Maura Borrego. Development of engineering education as a rigorous discipline: A study of the publication patterns of four coalitions. *Journal of Engineering Education*, 96(1):5–18, 2007.

[Boser *et al.* 2003] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 2003.

[Bowden *et al.* 2019] Kevin K Bowden, Shereen Oraby, Amita Misra, Jiaqi Wu, Stephanie Lukin, and Marilyn Walker. Data-driven dialogue systems for social agents. In *Advanced Social Interaction with Agents*, pages 53–56. Springer, 2019.

[Bower 2008] Matt Bower. A taxonomy of task types in computing. In *ACM SIGCSE Bulletin*, volume 40, pages 281–285. ACM, 2008.

[Boyd and Ellison 2007] Danah M Boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13:210–230, 2007.

[boyd 2004a] danah michele boyd. Friendster and publicly articulated social networking. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, page 1279â€“1282, New York, NY, USA, 2004. Association for Computing Machinery.

[Boyd 2004b] Danah Michele Boyd. Friendster and publicly articulated social networking. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1279–1282, 2004.

[boyd 2004c] danah michele boyd. Friendster and publicly articulated social networking. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, page 1279â€“1282, New York, NY, USA, 2004. Association for Computing Machinery.

[Bozkurt *et al.* 2007a] Ilker Nadi Bozkurt, Ozgur Baghoglu, and Erkan Uyar. Authorship attribution. In *2007 22nd international symposium on computer and information sciences*, pages 1–5. IEEE, 2007.

[Bozkurt *et al.* 2007b] Ilker Nadi Bozkurt, Ozgur Baghoglu, and Erkan Uyar. Authorship attribution. In *2007 22nd international symposium on computer and information sciences*, pages 1–5, Nov 2007.

[Bozzelli *et al.* 2015] Laura Bozzelli, Bastien Maubert, and Sophie Pinchinat. Uniform strategies, rational relations and jumping automata. *Information and Computation*, 242:80–107, 2015.

[Breiman 2001] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[Brennan and Greenstadt 2009] Michael Robert Brennan and Rachel Greenstadt. Practical attacks against authorship recognition techniques. In *Twenty-First IAAI Conference*, 2009.

[Brennan *et al.* 2012] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22, 2012.

[Bringhurst 2013] Robert Bringhurst. *The Elements of Typographic Style*. Version 4.0: 20th Anniversary Edition. Hartley & Marks Publishers, Point Roberts, WA, USA, 2013.

[Broadhurst and Trivedi 2018] Roderic Broadhurst and Harshit Trivedi. Malware in spam email: Trends in the 2016 australian spam intelligence data. *Available at SSRN 3413442*, 2018.

[Brocardo and Traore 2014] Marcelo Luiz Brocardo and Issa Traore. Continuous authentication using micro-messages. In *2014 Twelfth Annual International Conference on Privacy, Security and Trust*, pages 179–188, July 2014.

[Brocardo *et al.* 2017] Marcelo Luiz Brocardo, Issa Traore, Isaac Woungang, and Mohammad S. Obaidat. Authorship verification using deep belief network systems. *International Journal of Communication Systems*, 30(12):e3259, 2017. e3259 dac.3259.

[Brodie 2014] Karin Brodie. Learning about learner errors in professional learning communities. *Educational Studies in Mathematics*, 85(2):221–239, 2014.

[Brooks 1977] Ruven Brooks. Towards a theory of the cognitive processes in computer programming. *International Journal of Man-Machine Studies*, 9(6):737–751, 1977.

[Brown 1998] C Marlin Brown. *Human-computer interface design guidelines*. Intellect Books, 1998.

[Brown 2009] Gavin Brown. A new perspective for information theoretic feature selection. In *Artificial intelligence and statistics*, pages 49–56. PMLR, 2009.

[Brownlee and Ghiabi 2016] Billie Jeanne Brownlee and Maziyar Ghiabi. Passive, silent and revolutionary: The â€~arab springâ€™revisited. *Middle East Critique*, 25(3):299–316, 2016.

[Brüggemann-Klein and Wood 1998] Anne Brüggemann-Klein and Derick Wood. One-unambiguous regular languages. *Information and Computation*, 142(2):182–206, 1998.

[Bruner and others 1966] Jerome Seymour Bruner et al. *Toward a theory of instruction*, volume 59. Harvard University Press, 1966.

[Brusilovsky *et al.* 2008] Peter Brusilovsky, Sergey Sosnovsky, Danielle H Lee, Michael Yudelson, Vladimir Zadorozhny, and Xin Zhou. An open integrated exploratorium for database courses. In *ACM Special Interest Group on Computer Science Education Bulletin*, volume 40, pages 22–26. ACM, 2008.

[Bui and Zeng-Treitler 2014] Duy Duc An Bui and Qing Zeng-Treitler. Learning regular expressions for clinical text classification. *Journal of the American Medical Informatics Association*, 21(5):850–857, 2014.

[Burbules 2016] Nicholas C Burbules. How we use and are used by social media in education. *Educational Theory*, 66(4):551–565, 2016.

[Burileanu 2008] Dragos Burileanu. Spoken language interfaces for embedded applications. In *Human factors and voice interactive systems*, pages 135–161. Springer, 2008.

[Burke *et al.* 2010] Moira Burke, Cameron Marlow, and Thomas Lento. Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1909–1912, 2010.

[Burnap *et al.* 2015] Pete Burnap, Omer F Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 95:96–108, 2015.

[Burrows *et al.* 2009] Steven Burrows, Alexandra L. Uitdenbogerd, and Andrew Turpin. Application of information retrieval techniques for source code authorship attribution. In Xiaofang Zhou, Haruo Yokota, Ke Deng, and Qing Liu, editors, *Database Systems for Advanced Applications*, pages 699–713, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[Burrows 1987] John F Burrows. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary & Linguistic Computing*, 2(2):61–70, 1987.

[Burrows 1992] John F Burrows. Computers and the study of literature. *Computers and written texts*, pages 167–204, 1992.

[Burrows 2002] John Burrows. â€~Deltaâ€™: a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287, 09 2002.

[Burtsev *et al.* 2018] Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, et al. Deeppavlov: open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, 2018.

[Byers 2022] Kyle Byers. *How Many Blogs Are There? (And 141 Other Blogging Stats)*, 2022.

[Byrne 2007] Dara N Byrne. Public discourse, community concerns, and civic engagement: Exploring black social networking traditions on blackplanet. com. *Journal of computer-mediated communication*, 13(1):319–340, 2007.

[Cagliero *et al.* 2018] Luca Cagliero, Luigi De Russis, Laura Farinetti, and Teodoro Montanaro. Improving the effectiveness of SQL learning practice: a data-driven approach. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 980–989. IEEE, 2018.

[Cai *et al.* 2017] Ruichu Cai, Boyan Xu, Xiaoyan Yang, Zhenjie Zhang, Zijian Li, and Zhihao Liang. An encoder-decoder framework translating natural language to database queries. *arXiv preprint arXiv:1711.06061*, 2017.

[Caldeira 2008] Carlos Pampulim Caldeira. Teaching SQL: a case study. In *ACM Special Interest Group on Computer Science Education Bulletin*, volume 40, pages 340–340. ACM, 2008.

[Cambria and White 2014] Erik Cambria and Bebo White. Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

[Cambria *et al.* 2015] Erik Cambria, Jie Fu, Federica Bisio, and Soujanya Poria. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[Cambria 2016] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[Câmpeanu *et al.* 2003] Cezar Câmpeanu, Kai Salomaa, and Sheng Yu. A formal study of practical regular expressions. *International Journal of Foundations of Computer Science*, 14(06):1007–1018, 2003.

[Canales *et al.* 2011] Omar Canales, Vinnie Monaco, Thomas Murphy, Edyta Zych, John Stewart, Charles Tappert Alex Castro, Ola Sotoye, Linda Torres, and Greg Truley. A stylometry system for authenticating students taking online tests. *P. of Student-Faculty Research Day, Ed., CSIS. Pace University*, 2011.

[Cantador *et al.* 2011] Ivan Cantador, Peter L Brusilovsky, and Tsvi Kuflik. Second workshop on information heterogeneity and fusion in recommender systems (hetrec2011). 2011.

[Canter 1992] David Canter. An evaluation of the cusum stylistic analysis of confessions. *Expert Evidence*, 1(2):93–99, 1992.

[Cao *et al.* 2003] LJ Cao, Kok Seng Chua, WK Chong, HP Lee, and QM Gu. A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomputing*, 55(1-2):321–336, 2003.

[Cao *et al.* 2009] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9):1775–1781, 2009.

[Cappel 2002] James J Cappel. Entry-level is job skills: A survey of employers. *Journal of Computer Information Systems*, 42(2):76–82, 2002.

[Cappers and van Wijk 2017] Bram CM Cappers and Jarke J van Wijk. Exploring multivariate event sequences using rules, aggregations, and selections. *IEEE transactions on visualization and computer graphics*, 24(1):532–541, 2017.

[Cassidy 2004] Simon Cassidy. Learning styles: An overview of theories, models, and measures. *Educational psychology*, 24(4):419–444, 2004.

[Catarci *et al.* 1997] Tiziana Catarci, Maria F Costabile, Stefano Levialdi, and Carlo Batini. Visual query systems for databases: A survey. *Journal of Visual Languages & Computing*, 8(2):215–260, 1997.

[Ceballos *et al.* 2012] Ariel Gregorio Ceballos, Michael Anthony Caruso, and Kenneth Klaus. *Criteria builder for query builder*, November 13 2012. US Patent 8,312,038.

[Cebollero *et al.* 2015] Miguel Cebollero, Jay Natarajan, and Michael Coles. Error handling and dynamic SQL. In *Pro T-SQL Programmer's Guide*, pages 589–612. Springer, 2015.

[Cegielski and Jones-Farmer 2016] Casey G Cegielski and L Allison Jones-Farmer. Knowledge, skills, and abilities for entry-level business analytics positions: A multi-method study. *Decision Sciences Journal of Innovative Education*, 14(1):91–118, 2016.

[Cembalo *et al.* 2011] Maurizio Cembalo, Alfredo De Santis, and Umberto Ferraro Petrillo. SAVI: a new system for advanced SQL visualization. In *Proceedings of the 2011 conference on Information technology education*, pages 165–170. ACM, 2011.

[Cereda and Neto 2017] Paulo Roberto Massa Cereda and João José Neto. Instrumenting a context-free language recognizer. In *International Conference on Enterprise Information Systems*, pages 203–210, 2017.

[Cerra *et al.* 2014] Daniele Cerra, Mihai Datcu, and Peter Reinartz. Authorship analysis based on data compression. *Pattern Recognition Letters*, 42:79–84, 2014.

[Chaffey 2022] Dave Chaffey. *Global Social Media Statistics Research Summary 2022 [June 2022]*. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/, 2022.

[Chaiklin 2003] Seth Chaiklin. The zone of proximal development in Vygotskyâ€™s analysis of learning and instruction. *Vygotskyâ€™s educational theory in cultural context*, 1:39–64, 2003.

[Chakrabarti 2004] Soumen Chakrabarti. Breaking through the syntax barrier: Searching with entities and relations. In *European Conference on Machine Learning*, pages 9–16. Springer, 2004.

[Chamberlin 2012] Donald D Chamberlin. Early history of SQL. *IEEE Annals of the History of Computing*, 34(4):78–82, 2012.

[Chandarana *et al.* 2017] Dharmil Chandarana, Vraj Shah, Arun Kumar, and Lawrence Saul. SpeakQL: towards speech-driven multi-modal querying. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, page 11. ACM, 2017.

[Chandra and Suaib 2014] Ashish Chandra and Mohammad Suaib. A survey on web spam and spam 2.0. *International Journal of Advanced Computer Research*, 4(2):634, 2014.

[Chao 2016] Po-Yao Chao. Exploring students' computational practice, design and performance of problem-solving through a visual programming environment. *Computers & Education*, 95:202–215, 2016.

[Chaski 2001] Carole E Chaski. Empirical evaluations of language-based author identification techniques. *Forensic linguistics*, 8:1–65, 2001.

[Chaski 2005] Carole E Chaski. Whoâ€™s at the keyboard? authorship attribution in digital evidence investigations. *International journal of digital evidence*, 4(1):1–13, 2005.

[Chaski 2013] Carole E Chaski. Iler: A web-accessible resource for research in forensic linguistics. *Linguistic Evidence in Security, Law and Intelligence*, 1(1):99–103, 2013.

[Chaturvedi *et al.* 2018] Iti Chaturvedi, Erik Cambria, Roy E Welsch, and Francisco Herrera. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77, 2018.

[Chen and Chen 2015] Yu-Ren Chen and Hsin-Hsi Chen. Opinion spam detection in web forum: a real case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 173–183. International World Wide Web Conferences Steering Committee, 2015.

[Chen and Karahanna 2018] Adela Chen and Elena Karahanna. Life interrupted: The effects of technology-mediated work interruptions on work and nonwork outcomes. *MIS quarterly*, 42(4):1023–1042, 2018.

[Chen *et al.* 2016] Tao Chen, Ruifeng Xu, Yulan He, Yunqing Xia, and Xuan Wang. Learning user and product distributed representations using a sequence model for sentiment analysis. *IEEE Computational Intelligence Magazine*, 11(3):34–44, 2016.

[Chen *et al.* 2017a] Cen Chen, Kenli Li, Mingxing Duan, and Keqin Li. Extreme learning machine and its applications in big data processing. In *Big data analytics for sensor-network collected intelligence*, pages 117–150. Elsevier, 2017.

[Chen *et al.* 2017b] Xingliang Chen, Antonija Mitrovic, and Moffat Mathews. How much learning support should be provided to novices and advanced students? In *IEEE 17th International Conference on Advanced Learning Technologies*, pages 39–43. IEEE, 2017.

[Chen *et al.* 2017c] Yun-Nung Chen, Asli Celikyilmaz, and Dilek Hakkani-Tür. Deep learning for dialogue systems. In *Proceedings of ACL 2017, Tutorial Abstracts*, pages 8–14, 2017.

[Chen *et al.* 2018] Yuanyuan Chen, Yisheng Lv, Xiao Wang, Lingxi Li, and Fei-Yue Wang. Detecting traffic information from social media texts with deep learning approaches. *IEEE Transactions on Intelligent Transportation Systems*, 20(8):3049–3058, 2018.

[Chen *et al.* 2020] Jui-Lung Chen, Apritika Dermawan, et al. The influence of youtube beauty vloggers on indonesian consumersâ€™ purchase intention of local cosmetic products. *International Journal of Business and Management*, 15(5):100–116, 2020.

[Chen *et al.* 2022] Kaiping Chen, Zening Duan, and Sijia Yang. Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences*, 41(1):114–130, 2022.

[Chen *et al.* 2023] Mu-Yen Chen, Yi-Wei Lai, and Jiunn-Woei Lian. Using deep learning models to detect fake news about covid-19. *ACM Transactions on Internet Technology*, 23(2):1–23, 2023.

[Chen 2021] Beichen Chen. *Stylometric Embeddings for Book Similarities*, 2021.

[Cheng and Hüllermeier 2009] Weiwei Cheng and Eyke Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2):211–225, 2009.

[Cherpas 1992] Chris Cherpas. Natural language processing, pragmatics, and verbal behavior. *The Analysis of verbal behavior*, 10(1):135–147, 1992.

[Cheung *et al.* 2016] Alvin Cheung, Armando Solar-Lezama, et al. Computer-assisted query formulation. *Foundations and Trends® in Programming Languages*, 3(1):1–94, 2016.

[Chiang *et al.* 2012] Roger HL Chiang, Paulo Goes, and Edward A Stohr. Business intelligence and analytics education, and program development: A unique opportunity for the information systems discipline. *ACM Transactions on Management Information Systems (TMIS)*, 3(3):12, 2012.

[Chomsky and Lightfoot 2002] Noam Chomsky and David W Lightfoot. *Syntactic structures*. Walter de Gruyter, 2002.

[Chomsky 1956] Noam Chomsky. Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124, 1956.

[Chomsky 1959] Noam Chomsky. On certain formal properties of grammars. *Information and control*, 2(2):137–167, 1959.

[Chong *et al.* 2017] Frederic T Chong, Diana Franklin, and Margaret Martonosi. Programming languages and compiler design for realistic quantum hardware. *Nature*, 549(7671):180, 2017.

[Chou and Chou 2000] David C Chou and Amy Y Chou. A guide to the Internet revolution in banking. *Information Systems Management*, 17(2):51–57, 2000.

[Chouchane *et al.* 2013] Radhouane Chouchane, Natalia Stakhanova, Andrew Walenstein, and Arun Lakhotia. Detecting machine-morphed malware variants via engine attribution. *Journal of Computer Virology and Hacking Techniques*, 9(3):137–157, 2013.

[Chowdhury *et al.* 2022] Anjir Ahmed Chowdhury, Argho Das, Khadija Kubra Shahjalal Hoque, and Debajyoti Karmaker. A comparative study of hyperparameter optimization techniques for deep learning. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 509–521. Springer, 2022.

[Chowdhury 2003] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[Christina *et al.* 2010] V Christina, S Karpagavalli, and G Suganya. A study on email spam filtering techniques. *International Journal of Computer Applications*, 12(1):7–9, 2010.

[Chu *et al.* 2017] Shumo Chu, Chenglong Wang, Konstantin Weitz, and Alvin Cheung. Cosette: An automated prover for SQL. In *Conference on Innovative Data Systems Research*, 2017.

[Church and de Oliveira 2013] Karen Church and Rodrigo de Oliveira. What's up with Whatsapp?: comparing mobile instant messaging behaviors with traditional SMS. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services*, pages 352–361. ACM, 2013.

[Churches 2010] Andrew Churches. *Bloom's digital taxonomy*, 2010.

[Cienciala *et al.* 2014] Ludec Cienciala, Lucie Ciencialová, and Erzsébet Csuhaj Varjú. Towards p colonies processing strings. *Proceedings of the Twelfth Brainstorming Week on Membrane Computing, 103-118, Sevilla, ETS de Ingeniería Informática, 3-7 de Febrero, 2014,*, 2014.

[Claxton 2005] Larry D Claxton. Scientific authorship: Part 1. a window into scientific fraud? *Mutation Research/Reviews in Mutation Research*, 589(1):17–30, 2005.

[Clayer *et al.* 2013] Jean-Pierre Clayer, Claudine Toffolon, and Christophe Choquet. Patterns, pedagogical design schemes and process for instructional design. In *2013 IEEE 13th International Conference on Advanced Learning Technologies*, pages 304–306. IEEE, 2013.

[Cliche *et al.* 2017] Mathieu Cliche, David Rosenberg, Dhruv Madeka, and Connie Yee. Scatteract: Automated extraction of data from scatter plots. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part I 10*, pages 135–150. Springer, 2017.

[Click and Petit 2010] Amanda Click and Joan Petit. Social networking and web 2.0 in information literacy. *The International Information & Library Review*, 42(2):137–142, 2010.

[Codd 1970] Edgar F Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, 1970.

[Codd 1974] Edgar F Codd. *Seven steps to rendezvous with the casual user*. IBM Corporation, 1974.

[Coe 2007] Thomas S Coe. Using the Bloomberg professional system for finance classes. *Journal of Financial Education*, pages 48–62, 2007.

[Colby *et al.* 1971] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. Artificial paranoia. *Artificial Intelligence*, 2(1):1–25, 1971.

[Colby 1975] Kenneth Mark Colby. *Artificial paranoia: a computer simulation of paranoid process*. Pergamon Press, 1975.

[Collin 2003] Peter Collin. *Dictionary of banking and finance*. Psychology Press, London, 2003.

[Collins and Halverson 2018] Allan Collins and Richard Halverson. *Rethinking education in the age of technology: The digital revolution and schooling in America*. Teachers College Press, 2018.

[Collins *et al.* 2020] Botambu Collins, Dinh Tuyen Hoang, Ngoc Thanh Nguyen, and Dosam Hwang. Fake news types and detection models on social media a state-of-the-art survey. In *Asian Conference on Intelligent Information and Database Systems*, pages 562–573. Springer, 2020.

[Collis and Moonen 2008] Betty Collis and Jef Moonen. Web 2.0 tools and processes in higher education: Quality perspectives. *Educational Media International*, 45(2):93–106, 2008.

[Collobert and Weston 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[Collobert *et al.* 2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

[Connolly and Begg 2005] Thomas M Connolly and Carolyn E Begg. *Database systems: a practical approach to design, implementation, and management*. Pearson Education, 2005.

[Connolly and Begg 2006] Thomas M Connolly and Carolyn E Begg. A constructivist-based approach to teaching database analysis and design. *Journal of Information Systems Education*, 17(1):43, 2006.

[Conover *et al.* 2011a] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.

[Conover *et al.* 2011b] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 192–199. IEEE, 2011.

[Corbara *et al.* 2022] Silvia Corbara, Berta Chulvi, Paolo Rosso, and Alejandro Moreo. Rhythmic and psycholinguistic features for authorship tasks in the spanish parliament: Evaluation and analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, pages 79–92. Springer, 2022.

[Corbara *et al.* 2023] Silvia Corbara, Alejandro Moreo, and Fabrizio Sebastiani. Syllabic quantity patterns as rhythmic features for latin authorship attribution. *Journal of the Association for Information Science and Technology*, 74(1):128–141, 2023.

[CoreNLP 2016] Stanford CoreNLP. a suite of core NLP tools. *URL http://nlp. stanford. edu/software/corenlp. shtml (Last accessed: 2019-09-06)*, page 3, 2016.

[Coronel and Morris 2016] Carlos Coronel and Steven Morris. *Database systems: design, implementation, & management*. Cengage Learning, 2016.

[Corritore and Wiedenbeck 1991] Cynthia L Corritore and Susan Wiedenbeck. What do novices learn during program comprehension? *International Journal of Human-Computer Interaction*, 3(2):199–222, 1991.

[Costa-Jussa and Fonollosa 2016] Marta R Costa-Jussa and José AR Fonollosa. Character-based neural machine translation. *arXiv preprint arXiv:1603.00810*, 2016.

[Couderc and Ferrero 2015] Benoît Couderc and Jérémy Ferrero. fr2SQL: Interrogation de bases de données en français. In *22ème Traitement Automatique des Langues Naturelles*, pages 1–13, 2015.

[Coulthard *et al.* 2016] Malcolm Coulthard, Alison Johnson, and David Wright. *An introduction to forensic linguistics: Language in evidence*. Routledge, 2016.

[Coulthard 2014] Malcolm Coulthard. Whose text is it? on the linguistic investigation of authorship. In *Discourse and social life*, pages 270–287. Routledge, 2014.

[Covington and McFall 2010] Michael A Covington and Joe D McFall. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100, 2010.

[Covington *et al.* 1994] Michael A Covington, Barbara J Grosz, and Fernando CN Pereira. *Natural language processing for Prolog programmers*. Prentice hall Englewood Cliffs (NJ), 1994.

[Coyotl-Morales *et al.* 2006] Rosa María Coyotl-Morales, Luis Villaseñor-Pineda, Manuel Montes-y Gómez, and Paolo Rosso. Authorship attribution using word sequences. In José Francisco Martínez-Trinidad, Jesús Ariel Carrasco Ochoa, and Josef Kittler, editors, *Progress in Pattern Recognition, Image Analysis and Applications*, pages 844–853, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[Crotty *et al.* 2016] Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. The case for interactive data exploration accelerators (ideas). In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–6, 2016.

[Crowe *et al.* 2008] Alison Crowe, Clarissa Dirks, and Mary Pat Wenderoth. Biology in Bloom: implementing Bloom's taxonomy to enhance student learning in biology. *CBEâ€"Life Sciences Education*, 7(4):368–381, 2008.

[Cruse 2011] Alan Cruse. Meaning in language: An introduction to semantics and pragmatics. 2011.

[Crystal and Davy 2016] David Crystal and Derek Davy. *Investigating english style*. Routledge, 2016.

[Crystal 2011] David Crystal. *Internet linguistics: A student guide*. Routledge, 2011.

[Custódio and Paraboni 2019] José Eleandro Custódio and Ivandré Paraboni. An ensemble approach to cross-domain authorship attribution. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 201–212. Springer, 2019.

[Cutler *et al.* 2012] Adele Cutler, D Richard Cutler, and John R Stevens. Random forests. *Ensemble machine learning: Methods and applications*, pages 157–175, 2012.

[Cutler *et al.* 2020] Josh Cutler, Matt Dickenson, Josh Cutler, and Matt Dickenson. Application programming interfaces. *Computational Frameworks for Political and Social Research with Python*, pages 87–97, 2020.

[Daelemans 2013] Walter Daelemans. Explanation in computational stylometry. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 451–462. Springer, 2013.

[Dahl *et al.* 2012] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42, 2012.

[Daigle 2007] B. Daigle. Enterprise to enterprise instant messaging. *Google Patents, US Patent App. 11/300,981*, 2007.

[Daks and Clark 2016] Alon Daks and Aidan Clark. Unsupervised authorial clustering based on syntactic structure. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 114–118, 2016.

[Dale *et al.* 2000] Robert Dale, Hermann Moisl, and Harold Somers. *Handbook of natural language processing*. CRC Press, 2000.

[Dale 2019] Robert Dale. NLP commercialisation in the last 25 years. *Natural Language Engineering*, 25(3):419–426, 2019.

[Dam 2023] Jauron Gunther Dam. Ceo's tweets and firm stock returns: A case study of elon musk and tesla. 2023.

[Danaparamita and Gatterbauer 2011] Jonathan Danaparamita and Wolfgang Gatterbauer. QueryViz: helping users understand SQL queries and their patterns. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 558–561. ACM, 2011.

[Dandapat 2007] Sandipan Dandapat. Part of specch tagging and chunking with maximum entropy model. In *Proceedings of the IJCAI Workshop on Shallow Parsing for South Asian Languages*, pages 29–32, 2007.

[Dann *et al.* 2008] Wanda P Dann, Stephen Cooper, and Randy Pausch. *Learning to program with Alice*. Prentice Hall Press, 2008.

[Dann *et al.* 2011] Wanda P Dann, Stephen Cooper, and Randy Pausch. *Learning to Program with Alice (w/CD ROM)*. Prentice Hall Press, 2011.

[Davey 2017] Graham Davey. *Applications of conditioning theory*. Routledge, 2017.

[David *et al.* 2022] Walter David, Michelle King-Okoye, Gianluca Sensidoni, Alessandro Capone, Irene Mugambwa, Stanislava Kraynova, and Beatriz Garmendia-Doval. Operationalizing a medical intelligence platform for humanitarian security in protracted crises. In *Modelling and Simulation for Autonomous Systems: 8th International Conference, MESAS 2021, Virtual Event, October 13–14, 2021, Revised Selected Papers*, pages 397–416. Springer, 2022.

[Davis *et al.* 1952] KH Davis, R Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

[Davis 1997] Philip Davis. What computer skills do employees expect from recent college graduates? *THE Journal (Technological Horizons in Education)*, 25(2):74, 1997.

[De Boer and Badke-Schaub 2013] Robert J De Boer and Petra Badke-Schaub. Interaction between emotions and mental models in engineering and design activities. In *Emotional Engineering vol. 2*, pages 149–164. Springer, 2013.

[De Houwer *et al.* 2013] Jan De Houwer, Dermot Barnes-Holmes, and Agnes Moors. What is learning? on the nature and merits of a functional definition of learning. *Psychonomic bulletin & review*, 20(4):631–642, 2013.

[De Marneffe and Manning 2008] Marie-Catherine De Marneffe and Christopher D Manning. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8, 2008.

[De Morgan 1851] Augustus De Morgan. On a point connected with the dispute about the invention of fluxions. In *Abstracts of the Papers Communicated to the Royal Society of London*, number 5, pages 599–600. The Royal Society London, 1851.

[De Raadt *et al.* 2006] Michael De Raadt, Stijn Dekeyser, and Tien Yu Lee. Do students SQLify? improving learning outcomes with peer review and enhanced computer assisted assessment of querying skills. In *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006*, pages 101–108. ACM, 2006.

[De Raadt *et al.* 2007] Michael De Raadt, Stijn Dekeyser, and Tien Yu Lee. A system employing peer review and enhanced computer assisted assessment of querying skills. *Informatics in Education*, 6(1):163, 2007.

[de Silva 2017] NH Nisansa D de Silva. Relational databases and biomedical big data. *Bioinformatics in MicroRNA Research*, pages 69–81, 2017.

[de Vel *et al.* 2001a] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Rec.*, 30(4):55â€“64, dec 2001.

[De Vel *et al.* 2001b] Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. Multi-topic e-mail authorship attribution forensics. In *Proceedings ACM Conference on Computer Security-Workshop on Data Mining for Security Applications*, 2001.

[Dearden and Finlay 2006] Andy Dearden and Janet Finlay. Pattern languages in HCI: A critical review. *Human–computer interaction*, 21(1):49–102, 2006.

[Decker *et al.* 2012] Dyan Decker, Alexandre Blanc, John Loveland, and Mona Clayton. Data mining: Analysis of structured and unstructured information. *A Guide to Forensic Accounting Investigation*, pages 333–362, 2012.

[Dekeyser *et al.* 2007] Stijn Dekeyser, Michael de Raadt, and Tien Yu Lee. Computer assisted assessment of SQL query skills. In *Proceedings of the eighteenth conference on Australasian database-Volume 63*, pages 53–62. Australian Computer Society, Inc., 2007.

[Demirbas *et al.* 2010] Murat Demirbas, Murat Ali Bayir, Cuneyt Gurcan Akcora, Yavuz Selim Yilmaz, and Hakan Ferhatosmanoglu. Crowd-sourced sensing and collaboration using twitter. In *2010 IEEE International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM)*, pages 1–9. IEEE, 2010.

[Demsar *et al.* 2021] Vlad Demsar, Jan Brace-Govan, Gavin Jack, and Sean Sands. The social phenomenon of trolling: understanding the discourse and social practices of online provocation. *Journal of Marketing Management*, 37(11-12):1058–1090, 2021.

[Deng and Allahverdyan 2016] Weibing Deng and Armen E Allahverdyan. Stochastic model for phonemes uncovers an author-dependency of their usage. *PloS one*, 11(4):e0152561, 2016.

[Deng *et al.* 2013] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al. Recent advances in deep learning for speech research at microsoft. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8604–8608. IEEE, 2013.

[Deng *et al.* 2014] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[Deng *et al.* 2019a] Xuelian Deng, Yuqing Li, Jian Weng, and Jilian Zhang. Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78:3797–3816, 2019.

[Deng *et al.* 2019b] Xuelian Deng, Yuqing Li, Jian Weng, and Jilian Zhang. Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78:3797–3816, 2019.

[Devlin *et al.* 2018a] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Devlin *et al.* 2018b] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2018.

[Devouard 2011] Florence Devouard. 1.3 collective intelligence and business enterprise 2.0. *Common Knowledge: The Challenge of Transdisciplinarity*, page 31, 2011.

[Dhankar *et al.* 2021] Abhishek Dhankar, Hamman Samuel, Fahim Hassan, Nawshad Farruque, FranÃ§ois Bolduc, and Osmar ZaÃ¯ane. Analysis of covid-19 misinformation in social media using transfer learning. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 880–885, Nov 2021.

[Dhar *et al.* 2022] Ankita Dhar, Himadri Mukherjee, Shibaprasad Sen, Md Obaidullah Sk, Amitabha Biswas, Teresa Gonçalves, and Kaushik Roy. Author identification from literary articles with visual features: A case study with bangla documents. *Future Internet*, 14(10):272, 2022.

[Di Minin *et al.* 2021] Enrico Di Minin, Christoph Fink, Anna Hausmann, Jens Kremer, and Ritwik Kulkarni. How to address data privacy concerns when using social media data in conservation science. *Conservation Biology*, 35(2):437–446, 2021.

[Di Persio and Honchar 2016] Luca Di Persio and Oleksandr Honchar. Artificial neural networks approach to the forecast of stock market price movements. *International Journal of Economics and Management Systems*, 1, 2016.

[Diale *et al.* 2016a] Melvin Diale, Christiaan Van Der Walt, Turgay Celik, and Abiodun Modupe. Feature selection and support vector machine hyper-parameter optimisation for spam detection. In *2016 Pattern Recognition Association of South Africa and robotics and mechatronics international conference (PRASA-RobMech)*, pages 1–7. IEEE, 2016.

[Diale *et al.* 2016b] Melvin Diale, Christiaan Van Der Walt, Turgay Celik, and Abiodun Modupe. Feature selection and support vector machine hyper-parameter optimisation for spam detection. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–7. IEEE, 2016.

[Dicle and Dicle 2018] Mehmet F Dicle and Betul Dicle. Content analysis: frequency distribution of words. *The Stata Journal*, 18(2):379–386, 2018.

[Diederich *et al.* 2003] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. Authorship attribution with support vector machines. *Applied intelligence*, 19:109–123, 2003.

[Dietrich *et al.* 1997] Suzanne W Dietrich, Eric Eckert, and Kevin Piscator. WinRDBI: a windows-based relational database educational tool. In *ACM Special Interest Group on Computer Science Education Bulletin*, volume 29, pages 126–130. ACM, 1997.

[Digate *et al.* 2013] Charles J Digate, Christopher F Herot, Tonytip Ketudat, Alexis M Kopikis, and Daniel J Teven. System and method for immediate and delayed real-time communication activities using availability data from communication through an external instant messaging system. *US Patent 8,375,092*, 2013.

[Dimitriadou *et al.* 2016] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. Aide: an active learning-based approach for interactive data exploration. *IEEE Transactions on Knowledge and Data Engineering*, 28(11):2842–2856, 2016.

[Ding *et al.* 2014] Shifei Ding, Xinzheng Xu, and Ru Nie. Extreme learning machine and its applications. *Neural Computing and Applications*, 25:549–556, 2014.

[Ding *et al.* 2015] Steven HH Ding, Benjamin CM Fung, and Mourad Debbabi. A visualizable evidence-driven approach for authorship attribution. *ACM Transactions on Information and System Security (TISSEC)*, 17(3):1–30, 2015.

[Ding *et al.* 2017] Steven HH Ding, Benjamin CM Fung, Farkhund Iqbal, and William K Cheung. Learning stylometric representations for authorship analysis. *IEEE transactions on cybernetics*, 49(1):107–121, 2017.

[Dixon 2022] S Dixon. *Annual Twitter user growth rate in the United States from 2020 to 2025*. https://www.statista.com/statistics/238729/twitters-annual-growth-rate-in-the-us/, 8 2022. last checked: 05.09.2011.

[Do *et al.* 2014] Quan Do, Rajeev K Agrawal, Dhana Rao, and Venkat N Gudivada. Automatic generation of SQL queries. *American Society for Engineering Education, Paper ID*, 8958, 2014.

[Dohleman 2006] Bethany S Dohleman. Exploratory social network analysis with pajek. *Psychometrika*, 71(3):605, 2006.

[Dollinger 2010] Robert Dollinger. SQL lightweight tutoring module–semantic analysis of SQL queries based on XML representation and LINQ. In *EdMedia: World Conference on Educational Media and Technology*, pages 3323–3328. Association for the Advancement of Computing in Education, 2010.

[Dömösi *et al.* 2016] Pál Dömösi, Szilárd Zsolt Fazekas, and Masami Ito. On Chomsky Hierarchy of Palindromic Languages. *Acta Cybern.*, 22(3):703–713, 2016.

[Donais *et al.* 2013] Jonathan A Donais, Richard A Frost, Shane M Peelar, and Robert A Roddy. A system for the automated author attribution of text and instant messages. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1484–1485, 2013.

[Donath *et al.* 1999] Judith Donath, Karrie Karahalios, and Fernanda Viegas. Visualizing conversation. *Journal of computer-mediated communication*, 4(4):JCMC442, 1999.

[Done and Murphy 2018] Elizabeth J Done and Mike Murphy. The responsibilisation of teachers: a neoliberal solution to the problem of inclusion. *Discourse: Studies in the Cultural Politics of Education*, 39(1):142–155, 2018.

[Dubey *et al.* 2016] Mohnish Dubey, Sourish Dasgupta, Ankit Sharma, Konrad Höffner, and Jens Lehmann. Asknow: A framework for natural language query formalization in sparql. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29–June 2, 2016, Proceedings 13*, pages 300–316. Springer, 2016.

[Dubey *et al.* 2020] Souvik Dubey, Payel Biswas, Ritwik Ghosh, Subhankar Chatterjee, Mahua Jana Dubey, Subham Chatterjee, Durjoy Lahiri, and Carl J Lavie. Psychosocial impact of covid-19. *Diabetes & Metabolic Syndrome: clinical research & reviews*, 14(5):779–788, 2020.

[Duffy and Jonassen 2013] Thomas M Duffy and David H Jonassen. *Constructivism and the technology of instruction: A conversation*. Routledge, 2013.

[Duh 2018] Kevin Duh. *Bayesian Analysis in Natural Language Processing*, 2018.

[Dumais *et al.* 1998] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. 1998.

[Duncan 2015] Andrew Duncan. Demonstration of Bloomberg terminal and instant Bloomberg messaging. *Personal Communication*, 2015.

[Dunn and Dunn 1978] Rita Stafford Dunn and Kenneth J Dunn. *Teaching students through their individual learning styles: A practical approach*. Prentice Hall, 1978.

[Dunn *et al.* 2016] Jonathan Dunn, Shlomo Argamon, Amin Rasooli, and Geet Kumar. Profile-based authorship analysis. *Digital Scholarship in the Humanities*, 31(4):689–710, 2016.

[Eberl and Kaiser 2018] Maximilian Eberl and Stephan Kaiser. *Organizational Routines Meet Experimental Psycholo: The Role of Implicit Learning in the Modification of Organizational Routines*. Springer, 2018.

[Edelmann *et al.* 2020] Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A Bail. Computational social science and sociology. *Annual Review of Sociology*, 46:61–81, 2020.

[Eden *et al.* 2018] Amnon H Eden, Epameinondas Gasparis, Jon Nicholson, and Rick Kazman. Round-trip engineering with the two-tier programming toolkit. *Software Quality Journal*, 26(2):249–271, 2018.

[Eder 2013] Maciej Eder. Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182, 11 2013.

[Eder 2015] Maciej Eder. Does size matter? authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182, 2015.

[Ehiemua and Omoera 2015] Kingsley I Ehiemua and Osakue S Omoera. Social media platforms of reality drama: A study of selected facebook accounts. *Journal of African Media Studies*, 7(2):185–201, 2015.

[Eissen and Stein 2006] Sven Meyer zu Eissen and Benno Stein. Intrinsic plagiarism detection. In *European conference on information retrieval*, pages 565–569. Springer, 2006.

[El-Affendi 2018] Mohammed A El-Affendi. The generative power of Arabic morphology and implications: A case for pattern orientation in Arabic corpus annotation and a proposed pattern ontology. In *5th International Symposium on Data Mining Applications*, pages 36–45. Springer, 2018.

[El and Kassou 2014] Sara El Manar El and Ismail Kassou. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12), 2014.

[Elbaz *et al.* 2020] Oded Elbaz, Tova Milo, and Amit Somech. Incremental based top-k similarity search framework for interactive-data-analysis sessions. EDBT, 2020.

[Elder 2009] Marvin Elder. *Natural language database querying*, April 30 2009. US Patent App. 12/004,255.

[Ellis and Dix 2007] Geoffrey Ellis and Alan Dix. A taxonomy of clutter reduction for information visualisation. *IEEE transactions on visualization and computer graphics*, 13(6):1216–1223, 2007.

[Ellis and Mansmann 2010] Geoffrey Ellis and Florian Mansmann. Mastering the information age solving problems with visual analytics. In *Eurographics*, volume 2, page 5, 2010.

[Elmasri 2017] Ramez Elmasri. *Fundamentals of database systems*. 2017.

[ElSayed 2015] Khaled Nasser ElSayed. An Arabic natural language interface system for a database of the Holy Quran. *International Journal of Advanced Research in Artificial Intelligence*, 4(7):9–14, 2015.

[Elsom-Cook 1987] Mark Elsom-Cook. Intelligent computer-aided instruction research at the open university. cite report no. 10. 1987.

[Engeström and Sannino 2012] Yrjö Engeström and Annalisa Sannino. Whatever happened to process theories of learning? *Learning, Culture and Social Interaction*, 1(1):45–56, 2012.

[Entwistle and Ramsden 2015] Noel Entwistle and Paul Ramsden. *Understanding student learning (Routledge revivals)*. Routledge, 2015.

[Eriksson *et al.* 2005] Kent Eriksson, Katri Kerem, and Daniel Nilsson. Customer acceptance of Internet banking in Estonia. *International Journal of Bank Marketing*, 23(2):200–216, 2005.

[Eriksson 2014] Linda Eriksson. *Sequential Aggregation of Textual Features forDomain Independent Author Identication*, 2014.

[Ertmer and Newby 2013] Peggy A Ertmer and Timothy J Newby. Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance improvement quarterly*, 26(2):43–71, 2013.

[Escalante *et al.* 2011] Hugo Jair Escalante, Thamar Solorio, and Manuel Montes. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 288–298, 2011.

[Esmaeilzadeh and Taghva 2022] Armin Esmaeilzadeh and Kazem Taghva. Text classification using neural network language model (nnlm) and bert: An empirical comparison. In *Proceedings of SAI Intelligent Systems Conference*, pages 175–189. Springer, 2022.

[Etaywe 2018] Awni Shati Etaywe. A computerized analysis of gender linguistic patterns as reflected in jordaniansâ€™ facebook statuses: Lexical items, affect, theme, identity and more. *International Journal of Computer Applications*, 975:8887, 2018.

[Etzioni and Kautz 2002] Oren Etzioni and Henry Kautz. High precision natural language interfaces to databases: a graph theoretic approach draft. 2002.

[Evgeniou and Pontil 2001] Theodoros Evgeniou and Massimiliano Pontil. Workshop on support vector machines: theory and applications. 2001.

[Explosion 2017] AI Explosion. spaCy-industrial-strength natural language processing in Python. *URL: https://spacy. io*, 2017.

[Ezpeleta *et al.* 2016] Enaitz Ezpeleta, Urko Zurutuza, and José María Gómez Hidalgo. Does sentiment analysis help in Bayesian spam filtering? In *International Conference on Hybrid Artificial Intelligence Systems*, pages 79–90. Springer, 2016.

[Fabien *et al.* 2020a] Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India, December 2020. NLP Association of India (NLPAI).

[Fabien *et al.* 2020b] Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, 2020.

[Fagin *et al.* 2015] Ronald Fagin, Benny Kimelfeld, Frederick Reiss, and Stijn Vansummeren. Document spanners: A formal approach to information extraction. *Journal of the ACM (JACM)*, 62(2):12, 2015.

[Fang *et al.* 2005] Xiang Fang, Sooun Lee, and Seokha Koh. Transition of knowledge/skills requirement for entry-level is professionals: An exploratory study based on recruiters' perception. *Journal of Computer Information Systems*, 46(1):58–70, 2005.

[Farajian *et al.* 2017] M Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, 2017.

[Faroult and Robson 2006] Stephane Faroult and Peter Robson. *The art of SQL*. " O'Reilly Media, Inc.", 2006.

[Fast *et al.* 2016] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4647–4657, 2016.

[Fawcett 2006] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[Fedotova *et al.* 2021] Anastasia Fedotova, Aleksandr Romanov, Anna Kurtukova, and Alexander Shelupanov. Authorship attribution of social media and literary russian-language texts using machine learning methods and feature selection. *Future Internet*, 14(1):4, 2021.

[Felder *et al.* 1988] Richard M Felder, Linda K Silverman, et al. Learning and teaching styles in engineering education. *Engineering education*, 78(7):674–681, 1988.

[Felicia and Pitt 2009] Patrick Felicia and IJ Pitt. Profiling users in educational games. *Games-based learning advancement for multisensory human computer interfaces: Techniques and effective practices, Idea-Group, Hershey*, 2009.

[Feng *et al.* 2017] Huan Feng, Kassem Fawaz, and Kang G Shin. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, pages 343–355. ACM, 2017.

[Ferguson 2014] Mike Ferguson. E paper. *BMJ*, pages 72–73, 2014.

[Fernau *et al.* 2015] Henning Fernau, Meenakshi Paramasivan, and Markus L Schmid. Jumping finite automata: characterizations and complexity. In *International Conference on Implementation and Application of Automata*, pages 89–101. Springer, 2015.

[Ferracane *et al.* 2017] Elisa Ferracane, Su Wang, and Raymond Mooney. Leveraging discourse information effectively for authorship attribution. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 584–593, 2017.

[Fincher 1999] Sally Fincher. What are we doing when we teach programming? In *FIE'99 Frontiers in Education. 29th Annual Frontiers in Education Conference. Designing the Future of Science and Engineering Education. Conference Proceedings (IEEE Cat. No. 99CH37011*, volume 1, pages 12A4–1. IEEE, 1999.

[Fleming *et al.* 2011] Sandra Fleming, Gabrielle Mckee, and Sylvia Huntley-Moore. Undergraduate nursing students' learning styles: A longitudinal study. *Nurse education today*, 31(5):444–449, 2011.

[Fobbe 2020] Eilika Fobbe. Text-linguistic analysis in forensic authorship attribution. *JLL*, 9:93, 2020.

[Folland 2016] Kristin Annabel Torjussen Folland. *viSQLizer: An interactive visualizer for learning SQL*. Master's thesis, 2016.

[Foltỳnek *et al.* 2019] Tomáš Foltỳnek, Norman Meuschke, and Bela Gipp. Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR)*, 52(6):1–42, 2019.

[Forehand 2010] Mary Forehand. Bloomâ€™s taxonomy. *Emerging perspectives on learning, teaching, and technology*, 41(4):47–56, 2010.

[Forman and others 2003] George Forman et al. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3(Mar):1289–1305, 2003.

[Forsyth and Holmes 1996] RS Forsyth and DI Holmes. Feature-finding for text classification. *Literary and Linguistic Computing*, 11(4):163–174, 12 1996.

[Foster and Godbole 2016] Elvis C Foster and Shripad Godbole. Overview of SQL. In *Database Systems*, pages 205–209. Springer, 2016.

[Fourkioti *et al.* 2019] Olga Fourkioti, Symeon Symeonidis, and Avi Arampatzis. Language models and fusion for authorship attribution. *Information Processing & Management*, 56(6):102061, 2019.

[Frantzeskou *et al.* 2006a] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th international conference on Software engineering*, pages 893–896, 2006.

[Frantzeskou *et al.* 2006b] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. Source code author identification based on n-gram author profiles. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, *Artificial Intelligence Applications and Innovations*, pages 508–515, Boston, MA, 2006. Springer US.

[Frantzeskou *et al.* 2006c] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. Source code author identification based on n-gram author profiles. In Ilias Maglogiannis, Kostas Karpouzis, and Max Bramer, editors, *Artificial Intelligence Applications and Innovations*, pages 508–515, Boston, MA, 2006. Springer US.

[Frantzeskou *et al.* 2007] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, Carole E Chaski, and Blake Stephen Howald. Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *International Journal of Digital Evidence*, 6(1):1–18, 2007.

[Fraser 2015] Neil Fraser. Ten things we've learned from Blockly. In *2015 IEEE Blocks and Beyond Workshop (Blocks and Beyond)*, pages 49–50. IEEE, 2015.

[Fréry *et al.* 2014] Jordan Fréry, Christine Largeron, and Mihaela Juganaru-Mathieu. Ujm at clef in author identification–notebook for pan at clef 2014. *Cappellato, et al.[3]*, 2014.

[Freund and Schapire 1997] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[Friedl 2006] Jeffrey EF Friedl. *Mastering regular expressions*. "O'Reilly Media, Inc.", 2006.

[Fu *et al.* 2016] Rongrong Fu, Hong Wang, and Wenbo Zhao. Dynamic driver fatigue detection using hidden Markov model in real driving condition. *Expert Systems with Applications*, 63:397–411, 2016.

[Fu *et al.* 2018a] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Fu *et al.* 2018b] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[Fucks 1952] Wilhelm Fucks. On mathematical analysis of style. *Biometrika*, 39(1/2):122–129, 1952.

[Fung *et al.* 2020] Isaac Chun-Hai Fung, Elizabeth B Blankenship, Jennifer O Ahweyevu, Lacey K Cooper, Carmen H Duke, Stacy L Carswell, Ashley M Jackson, Jimmy C Jenkins III, Emily A Duncan, Hai Liang, et al. Public health implications of image-based social media: a systematic review of instagram, pinterest, tumblr, and flickr. *The Permanente Journal*, 24, 2020.

[Furey *et al.* 2000] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[Furst *et al.* 2002] Karen Furst, William W Lang, and Daniel E Nolle. Internet banking. *Journal of Financial Services Research*, 22(1-2):95–117, 2002.

[Gabarron *et al.* 2021] Elia Gabarron, Sunday Oluwafemi Oyeyemi, and Rolf Wynn. Covid-19-related misinformation on social media: a systematic review. *Bulletin of the World Health Organization*, 99(6):455, 2021.

[Galavotti *et al.* 2000] Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi. Experiments on the use of feature selection and negative evidence in automated text categorization. In *International Conference on Theory and Practice of Digital Libraries*, pages 59–68. Springer, 2000.

[Gale and Sampson 1995] William A Gale and Geoffrey Sampson. Good-turing frequency estimation without tears. *Journal of quantitative linguistics*, 2(3):217–237, 1995.

[Gales *et al.* 2008] Mark Gales, Steve Young, et al. The application of hidden Markov models in speech recognition. *Foundations and Trends® in Signal Processing*, 1(3):195–304, 2008.

[Gammon 2014] Anne Gammon. Over a quarter of americans have made malicious online comments. *YouGov, October*, 20:2014, 2014.

[Gamon 2004] Michael Gamon. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 611–617, 2004.

[Gantayat *et al.* 2019] Neelamadhav Gantayat, Diptikalyan Saha, Jaydeep Sen, and Senthil Mani. Goal-based ontology creation for natural language querying in sap-erp platform. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 231–237. ACM, 2019.

[Ganty and Valero 2019] Pierre Ganty and Pedro Valero. Regular expression search on compressed text. In *2019 Data Compression Conference (DCC)*, pages 528–537. IEEE, 2019.

[Gao *et al.* 2018] Chong-zhi Gao, Qiong Cheng, Pei He, Willy Susilo, and Jin Li. Privacy-preserving naive Bayes classifiers secure against the substitution-then-comparison attack. *Information Sciences*, 444:72–88, 2018.

[Gao *et al.* 2020a] Xin Gao, Jeno Szep, Pratik Satam, Salim Hariri, Sundaresh Ram, and Jeffrey J. RodrÃ-guez. Spatio-temporal processing for automatic vehicle detection in wide-area aerial video. *IEEE Access*, 8:199562–199572, 2020.

[Gao *et al.* 2020b] Xin Gao, Jeno Szep, Pratik Satam, Salim Hariri, Sundaresh Ram, and Jeffrey J Rodriguez. Spatio-temporal processing for automatic vehicle detection in wide-area aerial video. *IEEE Access*, 8:199562–199572, 2020.

[Garcia-Molina and Salem 1992] Hector Garcia-Molina and Kenneth Salem. Main memory database systems: An overview. *IEEE Transactions on Knowledge & Data Engineering*, (6):509–516, 1992.

[Gardner *et al.* 2018] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*, 2018.

[Garera and Yarowsky 2009] Nikesh Garera and David Yarowsky. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 710–718, 2009.

[Garner and Mariani 2015] Philip Garner and John Amedeo Mariani. Learning SQL in steps. *Journal on Systemics, Cybernetics and Informatics*, 13(4):19–24, 2015.

[Garner 2003] Stuart Garner. Learning resources and tools to aid novices learn programming. In *Informing science & information technology education joint conference (INSITE)*, pages 213–222, 2003.

[Gasparetto *et al.* 2022] Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. A survey on text classification algorithms: From text to predictions. *Information*, 13(2):83, 2022.

[Ge *et al.* 2016] Zhenhao Ge, Yufang Sun, and Mark Smith. Authorship attribution using a neural network language model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

[Geiger 2016] R Stuart Geiger. Bot-based collective blocklists in twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, 19(6):787–803, 2016.

[Gelbukh 2013] Alexander Gelbukh. *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*, volume 7816. Springer, 2013.

[Genise *et al.* 2019] Nicholas Genise, Craig Gentry, Shai Halevi, Baiyu Li, and Daniele Micciancio. Homomorphic encryption for finite automata. *IACR Cryptology ePrint Archive*, 2019:176, 2019.

[Gentner and Stevens 2014] Dedre Gentner and Albert L Stevens. *Mental models*. Psychology Press, 2014.

[Gers *et al.* 2000] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[Ghai *et al.* 2019] Ridhima Ghai, Sakshum Kumar, and Avinash Chandra Pandey. Spam detection using rating and review processing method. In *Smart Innovations in Communication and Computational Sciences*, pages 189–198. Springer, 2019.

[Gharavi *et al.* 2020] Erfaneh Gharavi, Neda Nazemi, and Faraz Dadgostari. *Early Outbreak Detection for Proactive Crisis Management Using Twitter Data: COVID-19 a Case Study in the US*, 2020.

[Gianfredi *et al.* 2018] Vincenza Gianfredi, Nicola Luigi Bragazzi, Daniele Nucci, Mariano Martini, Roberto Rosselli, Liliana Minelli, and Massimo Moretti. Harnessing big data for communicable tropical and sub-tropical disorders: implications from a systematic review of the literature. *Frontiers in Public Health*, 6:90, 2018.

[Gil and Lenz 2007] Joseph Yossi Gil and Keren Lenz. Simple and safe SQL queries with C++ templates. In *Proceedings of the 6th International Conference on Generative Programming and Component Engineering*, pages 13–24. ACM, 2007.

[Gill *et al.* 2017] Paul Gill, Emily Corner, Maura Conway, Amy Thornton, Mia Bloom, and John Horgan. Terrorist use of the internet by the numbers: Quantifying behaviors, patterns, and processes. *Criminology & Public Policy*, 16(1):99–117, 2017.

[Gill 2019] Navdeep Singh Gill. *Overview of artificial intelligence and natural language processing*, 2019.

[Gimpel *et al.* 2010] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. *Part-of-speech tagging for twitter: Annotation, features, and experiments*. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2010.

[Giordani 2008] Alessandra Giordani. Mapping natural language into SQL in a NLIDB. In *International Conference on Application of Natural Language to Information Systems*, pages 367–371. Springer, 2008.

[Glas *et al.* 2012] D Glas, Satoru Satake, Takayuki Kanda, and Norihiro Hagita. An interaction design framework for social robots. In *Robotics: Science and Systems*, volume 7, page 89, 2012.

[GligoriÄ‡ *et al.* 2018] Kristina GligoriÄ‡, Ashton Anderson, and Robert West. How constraints affect content: The case of twitterâ€™s switch from 140 to 280 characters. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), Jun. 2018.

[Gligorić *et al.* 2018] Kristina Gligorić, Ashton Anderson, and Robert West. How constraints affect content: The case of twitterâ€™s switch from 140 to 280 characters. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[Glorot and Bengio 2010a] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[Glorot and Bengio 2010b] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[Godinez and Jamil 2018] Josue Espinosa Godinez and Hasan M Jamil. Meet Cyrus-the Query by Voice mobile assistant for the tutoring and formative assessment of SQL learners. *arXiv preprint arXiv:1811.04160*, pages 1–6, 2018.

[Goel *et al.* 2019] Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. Hyst: A hybrid approach for flexible and accurate dialogue state tracking. *arXiv preprint arXiv:1907.00883*, 2019.

[Gogte *et al.* 2016] Vaibhav Gogte, Aasheesh Kolli, Michael J Cafarella, Loris D'Antoni, and Thomas F Wenisch. Hare: Hardware accelerator for regular expressions. In *The 49th Annual IEEE/ACM International Symposium on Microarchitecture*, page 44. IEEE Press, 2016.

[Golbeck and Hansen 2011] Jennifer Golbeck and Derek Hansen. Computing political preference among twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 1105â€"1108, New York, NY, USA, 2011. Association for Computing Machinery.

[Golbeck and Hansen 2014] Jennifer Golbeck and Derek Hansen. A method for computing political preference among twitter followers. *Social Networks*, 36:177–184, 2014. Special Issue on Political Networks.

[Gold and Sollich 2003] Carl Gold and Peter Sollich. Model selection for support vector machine classification. *Neurocomputing*, 55(1-2):221–249, 2003.

[Goldberg and Elhadad 2008] Yoav Goldberg and Michael Elhadad. splitSVM: Fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications. In *Proceedings of ACL-08: HLT, Short Papers*, pages 237–240, 2008.

[Goldberg and Levy 2014] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[Goldberg 2016] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[Gollub *et al.* 2013] Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Recent trends in digital text forensics and its evaluation. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 282–302, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[Golshaie 2019] Ramin Golshaie. Function words as idiolect markers: A corpus-based approach to authorship attribution in farsi. *Language Related Research*, 10(3):293–317, 2019.

[Gomes *et al.* 2017] Heitor M Gomes, Albert Bifet, Jesse Read, Jean Paul Barddal, Fabrício Enembreck, Bernhard Pfharinger, Geoff Holmes, and Talel Abdessalem. Adaptive random forests for evolving data stream classification. *Machine Learning*, 106:1469–1495, 2017.

[Gómez-Adorno *et al.* 2018a] Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and David Pinto. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100:741–756, 2018.

[Gómez-Adorno *et al.* 2018b] Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and David Pinto. Document embeddings learned on various types of n-grams for cross-topic authorship attribution. *Computing*, 100(7):741–756, 2018.

[González-Castro *et al.* 2013] Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. Class distribution estimation based on the hellinger distance. *Information Sciences*, 218:146–164, 2013.

[Gooch 2012] Phil Gooch. BADREX: In situ expansion and coreference of biomedical abbreviations using dynamic regular expressions. *arXiv preprint arXiv:1206.4522*, 2012.

[Goodfellow *et al.* 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[Goolsby 2009] Rebecca Goolsby. Lifting elephants: Twitter and blogging in global perspective. In *Social Computing and Behavioral Modeling*, pages 1–6, Boston, MA, 2009. Springer US.

[Gorman 2002] Michael E Gorman. Types of knowledge and their roles in technology transfer. *The Journal of Technology Transfer*, 27(3):219–231, 2002.

[Goyal *et al.* 2018] Palash Goyal, Sumit Pandey, and Karan Jain. Deep learning for natural language processing. *Apress, Berkeley*, 2018.

[Goyvaerts 2006] Jan Goyvaerts. *Regular Expressions: The Complete Tutorial*. Lulu Press, 2006.

[Grad 2012] Burton Grad. Relational database management systems: The formative years [guest editor's introduction]. *IEEE Annals of the History of Computing*, 34(4):7–8, 2012.

[Graesser *et al.* 2012] Arthur C Graesser, Mark W Conley, and Andrew Olney. Intelligent tutoring systems. 2012.

[Granneman 2010] Scott Granneman. Key utility tools for text & archives. In *Mac OS X Snow Leopard for Power Users*, pages 371–402. Springer, 2010.

[Grant 2020] Tim Grant. Text messaging forensics revisited. 2020.

[Graves and Schmidhuber 2005] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052. IEEE, 2005.

[Gray and Malins 2016] Carole Gray and Julian Malins. *Visualizing research: A guide to the research process in art and design*. Routledge, 2016.

[Gredler 1992] Margaret E Gredler. *Learning and instruction: Theory into practice*. Macmillan New York, 1992.

[Green Jr *et al.* 1961] Bert F Green Jr, Alice K Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pages 219–224. ACM, 1961.

[Greenhow and Li 2012] Christine Greenhow and Jiahang Li. Like, comment, share: Collaboration and civic engagement within social network sites. In *Emerging technologies for the classroom: A learning sciences perspective*, pages 127–141. Springer, 2012.

[Grefenstette *et al.* 2000] Gregory Grefenstette, Anne Schiller, and Salah Aït-Mokhtar. Recognizing lexical patterns in text. In *Lexicon Development for Speech and Language Processing*, pages 141–168. Springer, 2000.

[Greg 1944] WW Greg. *The statistical study of literary vocabulary*, 1944.

[Grieve 2007] Jack Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, 22(3):251–270, 2007.

[Griffiths and Steyvers 2004] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235, 2004.

[Grillenberger and Brinda 2012] Andreas Grillenberger and Torsten Brinda. eledSQL: a new web-based learning environment for teaching databases and SQL at secondary school level. In *Proceedings of the 7th Workshop in Primary and Secondary Computing Education*, pages 101–104. ACM, 2012.

[Grimmer 2010] Justin Ryan Grimmer. *Representational Style: The Central Role of Communication in Representation*. Harvard University, 2010.

[Grivas *et al.* 2015] Andreas Grivas, Anastasia Krithara, and George Giannakopoulos. Author profiling using stylometric and structural feature groupings. In *CLEF (Working Notes)*, 2015.

[Groff *et al.* 2002] James R Groff, Paul N Weinberg, et al. *SQL: the complete reference*, volume 2. McGraw-Hill/Osborne, 2002.

[Gross and Acquisti 2005] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, 2005.

[Große and Renkl 2007] Cornelia S Große and Alexander Renkl. Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, 17(6):612–634, 2007.

[Grover *et al.* 2015] Shuchi Grover, Roy Pea, and Stephen Cooper. Designing for deeper learning in a blended computer science course for middle school students. *Computer Science Education*, 25(2):199–237, 2015.

[Grune *et al.* 2012] Dick Grune, Kees Van Reeuwijk, Henri E Bal, Ceriel JH Jacobs, and Koen Langendoen. *Modern compiler design*. Springer Science & Business Media, 2012.

[Grust and Scholl 1999] Torsten Grust and Marc H Scholl. How to comprehend queries functionally. *Journal of Intelligent Information Systems*, 12(2-3):191–218, 1999.

[Gu *et al.* 2016] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016.

[Guillén-Nieto and Stein 2022] Victoria Guillén-Nieto and Dieter Stein. Introduction: Theory and practice in forensic linguistics. In *Language as Evidence*, pages 1–33. Springer, 2022.

[Guimaraes 2006] Mario Guimaraes. The kennesaw database courseware (kdc): strong points, weak points, and experience using it in a classroom environment. *Journal of Computing Sciences in Colleges*, 21(3):91–96, 2006.

[Gulwani *et al.* 2015] Sumit Gulwani, José Hernández-Orallo, Emanuel Kitzelmann, Stephen H Muggleton, Ute Schmid, and Benjamin Zorn. Inductive programming meets the real world. *Communications of the ACM*, 58(11):90–99, 2015.

[Gupta A. Band 2020] A Sharma S Bilakhiya K Gupta A. Band. *Detecting emotions behind the text*. https://shivamsharma26.github.io/text2emotion/, 2020.

[Gupta and others 2016] Deepa Gupta et al. Study on extrinsic text plagiarism detection techniques and tools. *Journal of Engineering Science & Technology Review*, 9(5), 2016.

[Gupta *et al.* 2017] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani, and B Tech. Study of twitter sentiment analysis using machine learning algorithms on python. *International Journal of Computer Applications*, 165(9):29–34, 2017.

[Gussenhoven and Jacobs 2017] Carlos Gussenhoven and Haike Jacobs. *Understanding phonology*. Routledge, 2017.

[Guthrie *et al.* 2007] David Guthrie, Louise Guthrie, Ben Allison, and Yorick Wilks. Unsupervised anomaly detection. In *IJCAI*, pages 1624–1628, 2007.

[Guthrie 2008] David Guthrie. *Unsupervised detection of anomalous text*. PhD thesis, University of Sheffield, 2008.

[Hadjadj and Sayoud 2021] Hassina Hadjadj and Halim Sayoud. Arabic authorship attribution using synthetic minority over-sampling technique and principal components analysis for imbalanced documents. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 15(4):1–17, 2021.

[Haghighi *et al.* 2018] Sepand Haghighi, Masoomeh Jasemi, Shaahin Hessabi, and Alireza Zolanvari. Pycm: Multiclass confusion matrix library in Python. *J. Open Source Software*, 3(25):729, 2018.

[Haiduc *et al.* 2010] S. Haiduc, J. Aponte, and A. Marcus. Supporting program comprehension with source code summarization. *ACM/IEEE 32nd International Conference on Software Engineering*, 2:223–226, 2010.

[Hale 2001] John Hale. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

[Halteren 2007] Hans Van Halteren. Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–17, 2007.

[Halvani *et al.* 2016] Oren Halvani, Christian Winter, and Anika Pflug. Authorship verification for different languages, genres and topics. *Digital Investigation*, 16:S33–S43, 2016.

[Halvani *et al.* 2020] Oren Halvani, Lukas Graner, and Roey Regev. Cross-domain authorship verification based on topic agnostic features. In *CLEF (Working Notes)*, 2020.

[Hamadi *et al.* 2022] Mostafa Hamadi, Jamal El-Den, Sami Azam, and Narumon Sriratanaviriyakul. Integrating social media as cooperative learning tool in higher education classrooms: An empirical study. *Journal of King Saud University-Computer and Information Sciences*, 34(6):3722–3731, 2022.

[Haman 2020] Michael Haman. The use of twitter by state leaders and its impact on the public during the covid-19 pandemic. *Heliyon*, 6(11), 2020.

[Hambrick 2007] Donald C Hambrick. The field of management's devotion to theory: Too much of a good thing? *Academy of management journal*, 50(6):1346–1352, 2007.

[Hamilton *et al.* 2018] James Hamilton, Manuel Gonzalez Berges, Jean-Charles Tournier, and Brad Schofield. Scada statistics monitoring using the elastic stack (elasticsearch, logstash, kibana). 2018.

[Han *et al.* 2013] Bo Han, Paul Cook, and Timothy Baldwin. unimelb: Spanish text normalisation. In *Tweet Normalisation Workshop at Sociedad Espanola para el Procesamiento del Lenguaje Natural (SEPLN)*, pages 32–36, 2013.

[Hancock 1986] Ian Hancock. The cryptolectal speech of the american roads: Traveler cant and american angloromani. *American Speech*, 61(3):206–220, 1986.

[Handler *et al.* 2016] Abram Handler, Matthew Denny, Hanna Wallach, and Brendan Oâ€™Connor. Bag of what? simple noun phrase extraction for text analysis. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 114–124, 2016.

[Hanneman and Riddle 2005] Robert A Hanneman and Mark Riddle. *Introduction to social network methods*, 2005.

[Hansen and Reich 2015] John D Hansen and Justin Reich. Democratizing education? examining access and usage patterns in massive open online courses. *Science*, 350(6265):1245–1248, 2015.

[Hansen *et al.* 2014] Niels Dalum Hansen, Christina Lioma, Birger Larsen, and Stephen Alstrup. Temporal context for authorship attribution. In *Information Retrieval Facility Conference*, pages 22–40. Springer, 2014.

[Hanus and Krone 2017] Michael Hanus and Julia Krone. A typeful integration of SQL into curry. *arXiv preprint arXiv:1701.00631*, 2017.

[Harasim 2017] Linda Harasim. *Learning theory and online technologies*. Routledge, 2017.

[Hardcastle 1997] RA Hardcastle. Cusum: a credible method for the determination of authorship? *Science & Justice: Journal of the Forensic Science Society*, 37(2):129–138, 1997.

[Harden 2017] Cynthia Harden. Regular expressions for IT men. 2017.

[Harrison 2015] Guy Harrison. Languages and programming interfaces. In *Next Generation Databases*, pages 167–190. Springer, 2015.

[Hashem *et al.* 2015] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of big data on cloud computing: Review and open research issues. *Information systems*, 47:98–115, 2015.

[Hastie *et al.* 2009] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[Hatami 2017] Zahra Hatami. Towards an understanding of the mental model process while writing SQL queries. 2017.

[Hayati *et al.* 2010] Pedram Hayati, Vidyasagar Potdar, Alex Talevski, Nazanin Firoozeh, Saeed Sarenche, and Elham A Yeganeh. Definition of spam 2.0: New spamming boom. In *4th IEEE International Conference on Digital Ecosystems and Technologies*, pages 580–584. IEEE, 2010.

[Haythornthwaite 2005] Caroline Haythornthwaite. Social networks and internet connectivity effects. *Information, Community & Society*, 8:125–147, 2005.

[Heimerl *et al.* 2014] Florian Heimerl, Steffen Lohmann, Simon Lange, and Thomas Ertl. Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii international conference on system sciences*, pages 1833–1842. IEEE, 2014.

[Hein 1999] George E Hein. The constructivist museum. *The educational role of the museum*, 2:73–79, 1999.

[Heinroth and Minker 2012] Tobias Heinroth and Wolfgang Minker. *Introducing spoken dialogue systems into Intelligent Environments*. Springer Science & Business Media, 2012.

[Heller 2019a] Jon Heller. Modify data with advanced DML. In *Pro Oracle SQL Development*, pages 191–218. Springer, 2019.

[Heller 2019b] Jon Heller. Understand relational databases. In *Pro Oracle SQL Development*, pages 3–28. Springer, 2019.

[Heller 2019c] Jon Heller. Write large SQL statements. In *Pro Oracle SQL Development*, pages 309–325. Springer, 2019.

[Hellweg 2011] Annie Hellweg. Social media sites of politicians influence their perception by constituents. *The Elon Journal of Undergraduate Research in Communications*, 2(1):22–36, 2011.

[Herring 2007] Susan C Herring. A faceted classification scheme for computer-mediated discourse. *Language@ internet*, 4(1), 2007.

[Herrmann *et al.* 2015] J Berenike Herrmann, Karina van Dalen-Oskam, and Christof Schöch. Revisiting style, a key concept in literary studies. *Journal of literary theory*, 9(1):25–52, 2015.

[Hess *et al.* 1989] Carla W Hess, Holly T Haug, and Richard G Landry. The reliability of type-token ratios for the oral language of school age children. *Journal of Speech, Language, and Hearing Research*, 32(3):536–540, 1989.

[Hewitt and Manning 2019] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[Hilton 1992] Ordway Hilton. *Scientific examination of questioned documents*. CRC press, 1992.

[Himdi *et al.* 2022] Hanen Himdi, George Weir, Fatmah Assiri, and Hassanin Al-Barhamtoshy. Arabic fake news detection based on textual analysis. *Arabian Journal for Science and Engineering*, pages 1–17, 2022.

[Hinh *et al.* 2016a] Robert Hinh, Sangmi Shin, and Julia Taylor. Using frame semantics in authorship attribution. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 004093–004098, Oct 2016.

[Hinh *et al.* 2016b] Robert Hinh, Sangmi Shin, and Julia Taylor. Using frame semantics in authorship attribution. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 004093–004098, 2016.

[Hinh *et al.* 2016c] Robert Hinh, Sangmi Shin, and Julia Taylor. Using frame semantics in authorship attribution. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 004093–004098. IEEE, 2016.

[Hinton *et al.* 2012a] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29:1–16, 2012.

[Hinton *et al.* 2012b] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[Hirschberg and Manning 2015] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[Hirschman and Gaizauskas 2001] Lynette Hirschman and Robert Gaizauskas. Natural language question answering: the view from here. *natural language engineering*, 7(4):275–300, 2001.

[Hirst and Feiguina 2007] Graeme Hirst and Ol'ga Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417, 2007.

[Hirtle 1996] Jeannine St Pierre Hirtle. Social constructivism. *English Journal*, 85(1):91, 1996.

[Hitschler *et al.* 2017] Julian Hitschler, Esther van den Berg, and Ines Rehbein. Authorship attribution with convolutional neural networks and pos-eliding. In *Proceedings of the Workshop on Stylistic Variation*, pages 53–58, 2017.

[Hmelo-Silver 2004] Cindy E Hmelo-Silver. Problem-based learning: What and how do students learn? *Educational psychology review*, 16(3):235–266, 2004.

[Ho 1995] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[Hoang *et al.* 2015] Duc Tam Hoang, Minh Le Nguyen, and Son Bao Pham. L2S: Transforming natural language questions into SQL queries. In *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pages 85–90. IEEE, 2015.

[Hoffman and Novak 2012] Donna L Hoffman and Thomas P Novak. *Toward a deeper understanding of social media*, 2012.

[Hoffman *et al.* 2010] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23, 2010.

[Hogan 2018] Rex Hogan. *A practical guide to database design*. Chapman and Hall/CRC, 2018.

[Hohman *et al.* 2020] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[Hoic-Bozic *et al.* 2009] Natasa Hoic-Bozic, Vedran Mornar, and Ivica Boticki. A blended learning approach to course design and implementation. *IEEE Transactions on Education*, 52(1):19–30, 2009.

[Holmes and Forsyth 1995] David I Holmes and Richard S Forsyth. The federalist revisited: New directions in authorship attribution. *Literary and Linguistic computing*, 10(2):111–127, 1995.

[Holmes and Tweedie 1995] David I Holmes and Fiona J Tweedie. Forensic stylometry: A review of the cusum controversy. *Revue Informatique et Statistique dans les Sciences Humaines*, 31(1):19–47, 1995.

[Holmes 1985] David I Holmes. The analysis of literary styleâ€"a review. *Journal of the Royal Statistical Society: Series A (General)*, 148(4):328–341, 1985.

[Holmes 1994a] David I Holmes. Authorship attribution. *Computers and the Humanities*, 28(2):87–106, 1994.

[Holmes 1994b] David I Holmes. Authorship attribution. *Computers and the Humanities*, 28(2):87–106, 1994.

[Holmes 1998] David I Holmes. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3):111–117, 1998.

[Holmes 2003] David I Holmes. Stylometry and the civil war: The case of the pickett letters. *Chance*, 16(2):18–25, 2003.

[Homem and Carvalho 2011] Nuno Homem and Joao Paulo Carvalho. Authorship identification and author fuzzy fingerprints. In *2011 Annual Meeting of the North American Fuzzy Information Processing Society*, pages 1–6. IEEE, 2011.

[Honey *et al.* 1992] Peter Honey, Alan Mumford, et al. The manual of learning styles. 1992.

[Hoover 2003] David L Hoover. Another Perspective on Vocabulary Richness. *Computers and the Humanities*, 37(2):151–178, 2003.

[Hoover 2004] David L Hoover. Testing burrows's delta. *Literary and linguistic computing*, 19(4):453–475, 2004.

[Horn *et al.* 2017] Ilana Seidel Horn, Brette Garner, Britnie Delinger Kane, and Jason Brasel. A taxonomy of instructional learning opportunities in teachersâ€™ workgroup conversations. *Journal of Teacher Education*, 68(1):41–54, 2017.

[Hosseinia and Mukherjee] Marjan Hosseinia and Arjun Mukherjee. A parallel hierarchical attention network for style change detection.

[Hosu *et al.* 2018] Ionel Alexandru Hosu, Radu Cristian Alexandru Iacob, Florin Brad, Stefan Ruseti, and Traian Rebedea. Natural language interface for databases using a Dual-Encoder model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 514–524, 2018.

[Hourrane and others 2019] Oumaima Hourrane et al. Rich style embedding for intrinsic plagiarism detection. *International Journal of Advanced Computer Science and Applications*, 10(11), 2019.

[Houvardas and Stamatatos 2006] John Houvardas and Efstathios Stamatatos. N-gram feature selection for authorship identification. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 77–86. Springer, 2006.

[Howedi and Mohd 2014] Fatma Howedi and Masnizah Mohd. Text classification for authorship attribution using naive bayes classifier with limited training data. *Computer Engineering and Intelligent Systems*, 5(4):48–56, 2014.

[Hoy 2018] Matthew B Hoy. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[Hu and Shepherd 2013] Helen H Hu and Tricia D Shepherd. Using POGIL to help students learn to program. *ACM Transactions on Computing Education (TOCE)*, 13(3):13, 2013.

[Hu *et al.* 2007] Ying Hu, Seema Sundara, and Jagannathan Srinivasan. Supporting time-constrained SQL queries in Oracle. In *Proceedings of the 33rd International Conference on very Large Databases*, pages 1207–1218. VLDB Endowment, 2007.

[Hu *et al.* 2013] Yuheng Hu, Kartik Talamadupula, and Subbarao Kambhampati. Dude, srsly?: The surprisingly formal nature of twitter's language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 244–253, 2013.

[Hu *et al.* 2016a] Baotian Hu, Buzhou Tang, Qingcai Chen, and Longbiao Kang. A novel word embedding learning model using the dissociation between nouns and verbs. *Neurocomputing*, 171:1108–1117, 2016.

[Hu *et al.* 2016b] Tianran Hu, Haoyuan Xiao, Jiebo Luo, and Thuy-vy Thi Nguyen. What the language you tweet says about your occupation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 181–190, 2016.

[Hu *et al.* 2020a] Zhiqiang Hu, Roy Ka-Wei Lee, Lei Wang, Ee-Peng Lim, and Bo Dai. Deepstyle: User style embedding for authorship attribution of short texts. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 221–229. Springer, 2020.

[Hu *et al.* 2020b] Zhiqiang Hu, Roy Ka-Wei Lee, Lei Wang, Ee-peng Lim, and Bo Dai. Deepstyle: User style embedding for authorship attribution of short texts. In Xin Wang, Rui Zhang, Young-Koo Lee, Le Sun, and Yang-Sae Moon, editors, *Web and Big Data*, pages 221–229, Cham, 2020. Springer International Publishing.

[Huang *et al.* 2012] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, 2012.

[Huang *et al.* 2014a] Chaochao Huang, Xipeng Qiu, and Xuanjing Huang. Text classification with document embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 131–140. Springer, 2014.

[Huang *et al.* 2014b] Faliang Huang, Chaoxiong Li, and Li Lin. Identifying gender of microblog users based on message mining. In *International Conference on Web-Age Information Management*, pages 488–493. Springer, 2014.

[Huang *et al.* 2014c] James Huang, Stephanie Rogers, and Eunkwang Joo. Improving restaurants by extracting subtopics from Yelp reviews. *iConference 2014 (Social Media Expo)*, 2014.

[Huang *et al.* 2014d] Xuedong Huang, James Baker, and Raj Reddy. A historical perspective of speech recognition. *Communications of the ACM*, 57(1):94–103, 2014.

[Huang *et al.* 2015a] Gao Huang, Guang-Bin Huang, Shiji Song, and Keyou You. Trends in extreme learning machines: A review. *Neural Networks*, 61:32–48, 2015.

[Huang *et al.* 2015b] Ting-Hao Kenneth Huang, Walter S Lasecki, and Jeffrey P Bigham. Guardian: A crowd-powered spoken dialog system for web apis. In *Third AAAI conference on human computation and crowdsourcing*, 2015.

[Hubbard 1995] EH Hubbard. Linguistic fingerprinting? a case study in forensic stylometrics. *South African Journal of Linguistics*, 13(sup26):55–72, 1995.

[Hughes *et al.* 2017] Mark Hughes, I Li, Spyros Kotoulas, and Toyotaro Suzumura. Medical text classification using convolutional neural networks. *Stud Health Technol Inform*, 235:246–250, 2017.

[Humphreys 2017] Ashlee Humphreys. Social media. In *The Routledge Companion to Consumer Behavior*, pages 363–379. Routledge, 2017.

[Hung 2001] David Hung. Theories of learning and computer-mediated instructional technologies. *Educational Media International*, 38(4):281–287, 2001.

[Hürlimann *et al.* 2015] Manuela Hürlimann, Benno Weck, Esther van den Berg, Simon Suster, and Malvina Nissim. Glad: Groningen lightweight authorship detection. In *CLEF (Working Notes)*, 2015.

[Hurst *et al.* 2013] Beth Hurst, Randall Wallace, and Sarah B Nixon. The impact of social interaction on student learning. *Reading Horizons*, 52(4):5, 2013.

[Hutchins and Somers 1992] William John Hutchins and Harold L Somers. *An introduction to machine translation*, volume 362. Academic Press London, 1992.

[Hutchins 1986] William John Hutchins. *Machine translation: past, present, future*. Ellis Horwood Chichester, 1986.

[Hutto and Gilbert 2014] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[Idris *et al.* 2015] Ismaila Idris, Ali Selamat, Ngoc Thanh Nguyen, Sigeru Omatu, Ondrej Krejcar, Kamil Kuca, and Marek Penhaker. A combined negative selection algorithm–particle swarm optimization for an email spam detection system. *Engineering Applications of Artificial Intelligence*, 39:33–44, 2015.

[Iedemska *et al.* 2014] Jane Iedemska, Gianluca Stringhini, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. The tricks of the trade: What makes spam campaigns successful? In *2014 IEEE Security and Privacy Workshops*, pages 77–83. IEEE, 2014.

[Inge 2013] Andre Inge. Hidden Markov models. 2013.

[Iqbal *et al.* 2008a] Farkhund Iqbal, Rachid Hadjidj, Benjamin C M Fung, and Mourad Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *digital investigation*, 5:S42–S51, 2008.

[Iqbal *et al.* 2008b] Farkhund Iqbal, Rachid Hadjidj, Benjamin CM Fung, and Mourad Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *digital investigation*, 5:S42–S51, 2008.

[Iqbal *et al.* 2013] Farkhund Iqbal, Hamad Binsalleeh, Benjamin CM Fung, and Mourad Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231:98–112, 2013.

[Ireland 2015] Katherine Ireland. The Bloomberg terminal: Mission critical or a "very expensive facebook"? *Harvard Business School Open Forum, Harvard Business School*, 2015.

[Isaac 2017] M Isaac. Twitter to test doubling tweet length to 280 characters. *The New York Times*, 26, 2017.

[Islam *et al.* 2019] Md Zahidul Islam, Jixue Liu, Jiuyong Li, Lin Liu, and Wei Kang. A semantics aware random forest for text classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1061–1070, 2019.

[Ismail and Homsi 2018] Walaa Saber Ismail and Masun Nabhan Homsi. DAWQAS: A dataset for Arabic why question answering system. *Procedia computer science*, 142:123–131, 2018.

[Jackson and Moulinier 2007] Peter Jackson and Isabelle Moulinier. *Natural language processing for online applications*. John Benjamins, 2007.

[Jacobs 2014] Paul S Jacobs. *Text-based intelligent systems: Current research and practice in information extraction and retrieval*. Psychology Press, 2014.

[Jafariakinabad and Hua 2019] Fereshteh Jafariakinabad and Kien A. Hua. Style-aware neural model with application in authorship attribution. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 325–328, 2019.

[Jafariakinabad *et al.* 2019] Fereshteh Jafariakinabad, Sansiri Tarnpradab, and Kien A. Hua. *Syntactic Recurrent Neural Network for Authorship Attribution*, 2019.

[Jäger and Rogers 2012] Gerhard Jäger and James Rogers. Formal language theory: refining the Chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1598):1956–1970, 2012.

[James 2006] Mary James. Assessment, teaching and theories of learning. *Assessment and learning*, 47:60, 2006.

[Jamil 2017] Hasan M Jamil. Knowledge rich natural language queries over structured biological databases. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 352–361. ACM, 2017.

[Jangid *et al.* 2018] Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, and Roger Zimmermann. Aspect-based financial sentiment analysis using deep learning. In *Companion Proceedings of the The Web Conference 2018*, pages 1961–1966. International World Wide Web Conferences Steering Committee, 2018.

[Jankowska *et al.* 2013] Magdalena Jankowska, Vlado Keselj, and Evangelos Milios. Proximity based one-class classification with common n-gram dissimilarity for authorship verification task. In *CLEF 2013 Evaluation Labs and Workshop–Working Notes Papers*, pages 23–26, 2013.

[Jankowska *et al.* 2014] Magdalena Jankowska, Evangelos Milios, and Vlado Keselj. Author verification using common n-gram profiles of text documents. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 387–397, 2014.

[Jansen *et al.* 2009] Bernard J Jansen, Danielle Booth, and Brian Smith. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management*, 45(6):643–663, 2009.

[Java *et al.* 2007] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: An analysis of a microblogging community. In *International Workshop on Social Network Mining and Analysis*, pages 118–138. Springer, 2007.

[Javed 2009] Faizan Javed. *Techniques for Context-Free Grammar Induction and Applications: Application of novel inference algorithms to software maintenance problems*. VDM Verlag, 2009.

[Jebb *et al.* 2017] Andrew T Jebb, Scott Parrigon, and Sang Eun Woo. Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27(2):265–276, 2017.

[Jeffrey 2016] Robin Jeffrey. *About Writing: A Guide*. 9 2016.

[Jenkins 2002] Tony Jenkins. On the difficulty of learning to program. In *Proceedings of the 3rd Annual Conference of the LTSN Centre for Information and Computer Sciences*, volume 4, pages 53–58. Citeseer, 2002.

[Jhamtani *et al.* 2017] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*, 2017.

[Jin and Murakami 2007] Mingzhe Jin and Masakatsu Murakami. Authorship identification using random forests. *Proceedings of the Institute of Statistical Mathematics*, 55(2):255–268, 2007.

[Jin 2014] Mingzhe Jin. Using integrated classification algorithm to identify a textâ€™s author. *The Japanese Journal of Behaviormetrics*, 41(1):35–46, 2014.

[Jindal and Liu 2008] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 219â€“230, New York, NY, USA, 2008. Association for Computing Machinery.

[Jing *et al.* 2019] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 2019.

[Joachims 1998] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[Jockers and Witten 2010] Matthew L Jockers and Daniela M Witten. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2):215–223, 2010.

[Johannsen *et al.* 2015] Anders Johannsen, Dirk Hovy, and Anders Søgaard. Cross-lingual syntactic variation over age and gender. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 103–112, 2015.

[Johnson and Fuller 2006] Colin G Johnson and Ursula Fuller. Is Bloom's taxonomy appropriate for computer science? In *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006*, pages 120–123. ACM, 2006.

[Johnson and others 2010] A Johnson et al. Forensic stylistics. theory and practice of forensic stylistics. 2010.

[Johnson and Sinanovic 2001] Don Johnson and Sinan Sinanovic. Symmetrizing the kullback-leibler distance. *IEEE Transactions on Information Theory*, 2001.

[Johnson and Soloway 1985] W Lewis Johnson and Elliot Soloway. PROUST: Knowledge-based program understanding. *IEEE Transactions on Software Engineering*, (3):267–275, 1985.

[Johnson-Laird 2010] Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

[Jolliffe and Cadima 2016] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[Jones *et al.* 2016] Kiku Jones, Lori NK Leonard, and Guido Lang. Desired skills for entry level IS positions: Identification and assessment. *Journal of Computer Information Systems*, pages 1–7, 2016.

[Jones *et al.* 2022] Keenan Jones, Jason RC Nurse, and Shujun Li. Are you robert or roberta? deceiving online authorship attribution models using neural text generators. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 429–440, 2022.

[Jones 1972] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[Joshi and Zincir-Heywood 2019] Mihir Joshi and Nur Zincir-Heywood. Classification of micro-texts using sub-word embeddings. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 526–533, 2019.

[Jost *et al.* 2005] Timothée Jost, Nabil Ouerhani, Roman Von Wartburg, René Müri, and Heinz Hügli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107–123, 2005.

[Jost *et al.* 2018] John T Jost, Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling, and Joshua A Tucker. How social media facilitates political protest: Information, motivation, and social networks. *Political psychology*, 39:85–118, 2018.

[Joulin *et al.* 2017] Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, 2017.

[Juang and Rabiner 1991] Biing Hwang Juang and Laurence R Rabiner. Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[Julavanich *et al.* 2019] Thanakrit Julavanich, Srinual Nalintippayawong, and Kanokwan Atchariyachanvanich. RSQLg: The reverse SQL question generation algorithm. In *2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA)*, pages 908–912. IEEE, 2019.

[Jungbluth *et al.* 2018] Jan Jungbluth, Rolf Krieger, Wolfgang Gerke, and Peter Plapper. Combining virtual and robot assistants - a case study about integrating Amazon's Alexa as a voice interface in robotics. In *Robotix-Academy Conference for Industrial Robotics*, page 5. Shaker, 2018.

[Juola and others 2008] Patrick Juola et al. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1(3):233–334, 2008.

[Juola and Vescovi 2010] Patrick Juola and Darren Vescovi. Empirical evaluation of authorship obfuscation using jgaap. In *Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security*, pages 14–18, 2010.

[Juola *et al.* 2006] Patrick Juola, John Sofko, and Patrick Brennan. A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21(2):169–178, 2006.

[Juola *et al.* 2019] Patrick Juola, George K Mikros, and Sean Vinsick. Correlations and potential cross-linguistic indicators of writing style. *Journal of Quantitative Linguistics*, 26(2):146–171, 2019.

[Juola 2006] Patrick Juola. Authorship attribution for electronic documents. In *IFIP International Conference on Digital Forensics*, pages 119–130. Springer, 2006.

[Juola 2007] Patrick Juola. Future trends in authorship attribution. In *IFIP International Conference on Digital Forensics*, pages 119–132. Springer, 2007.

[Juola 2010] Patrick Juola. Authorship attribution and the digital humanities curriculum. pages 1–21, 2010.

[Juola 2012] Patrick Juola. Large-scale experiments in authorship attribution. *English Studies*, 93(3):275–283, 2012.

[Juola 2013] Patrick Juola. How a computer program helped show jk rowling write a cuckooâ€™s calling. *Scientific American, August 20th*, 2013.

[Juola 2015] Patrick Juola. The rowling case: a proposed standard analytic protocol for authorship questions. *Digital Scholarship in the Humanities*, 30(suppl_1):i100–i113, 2015.

[Juola 2021] Patrick Juola. Verifying authorship for forensic purposes: A computational protocol and its validation. *Forensic Science International*, 325:110824, 2021.

[Jurafsky and Manning 2012] Dan Jurafsky and Christopher Manning. Natural language processing. *Instructor*, 212(998):3482, 2012.

[Jurafsky *et al.* 2014] Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, 19(4), Mar. 2014.

[Jurafsky 2000] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[Jusoh and Alfawareh 2012] Shaidah Jusoh and Hejab M Alfawareh. Techniques, applications and challenging issue in text mining. *International Journal of Computer Science Issues (IJCSI)*, 9(6):431, 2012.

[Kadiyala and Patgiri 2023] VS Nageswara Rao Kadiyala and Ripon Patgiri. An investigation on socio-cyber crime graph. In *Advances in Computers*, volume 128, pages 423–443. Elsevier, 2023.

[Kalchbrenner *et al.* 2014] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[Kalgutkar *et al.* 2019] Vaibhavi Kalgutkar, Ratinder Kaur, Hugo Gonzalez, Natalia Stakhanova, and Alina Matyukhina. Code authorship attribution: Methods and challenges. *ACM Computing Surveys (CSUR)*, 52(1):1–36, 2019.

[Kalina and Powell 2009] Cody Kalina and KC Powell. Cognitive and social constructivism: Developing tools for an effective classroom. *Education*, 130(2):241–250, 2009.

[Kamat *et al.* 2014] Niranjan Kamat, Prasanth Jayachandran, Karthik Tunga, and Arnab Nandi. Distributed and interactive cube exploration. In *2014 IEEE 30th international conference on data engineering*, pages 472–483. IEEE, 2014.

[Kamberg 2012] Mary-Lane Kamberg. *Evan Williams, Biz Stone, Jack Dorsey, and Twitter*. The Rosen Publishing Group, Inc, 2012.

[Kantorowitz 2014] Eliezer Kantorowitz. Verbal use case specifications for informal requirements elicitation. In *Building Bridges: HCI, Visualization, and Non-formal Modeling*, pages 165–174. Springer, 2014.

[Kapetanios 2008] Epaminondas Kapetanios. *Natural Language and Information Systems: 13th International Conference on Applications of Natural Language to Information Systems, NLDB 2008 London, UK, June 24-27, 2008, Proceedings*, volume 5039. Springer Science & Business Media, 2008.

[Kaplan and Mazurek 2018] Andreas Kaplan and Grzegorz Mazurek. Social media. *Handbook of media management and economics*, pages 273–286, 2018.

[Kapur *et al.* 2018] Arnav Kapur, Shreyas Kapur, and Pattie Maes. Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*, pages 43–53. ACM, 2018.

[Kapusta *et al.* 2021] Jozef Kapusta, Martin Drlik, and Michal Munk. Using of n-grams from morphological tags for fake news classification. *PeerJ Computer Science*, 7:e624, 2021.

[Karhumäki 2007] Juhani Karhumäki. *Automata and formal languages*, 2007.

[Kari 2013] Jarkko Kari. Automata and formal languages. *Fall semester*, 2013.

[Karlgren and Eriksson 2007] Jussi Karlgren and Gunnar Eriksson. Authors, genre, and linguistic convention. In *SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, 27 August 2007, Amsterdam*, 2007.

[Karlgren *et al.* 2018] Jussi Karlgren, Lewis Esposito, Chantal Gratton, and Pentti Kanerva. Authorship profiling without using topical information: Notebook for pan at clef 2018. In *CLEF (Working Notes)*, 2018.

[Kasavana *et al.* 2010] Michael L Kasavana, Khaldoon Nusair, and Katherine Teodosic. Online social networking: redefining the human web. *Journal of Hospitality and Tourism Technology*, 1(1):68–82, 2010.

[Katz and Shmallo 2016] Adi Katz and Ronit Shmallo. Learning from errors as a pedagogic approach for reaching a higher conceptual level in database modeling. In *International Conference on Advanced Information Systems Engineering*, pages 93–102. Springer, 2016.

[Kaur *et al.* 2019] Ravneet Kaur, Sarbjeet Singh, and Harish Kumar. Authorship analysis of online social media content. In *Proceedings of 2nd International Conference on Communication, Computing and Networking*, pages 539–549. Springer, 2019.

[Kearns *et al.* 1997] R Kearns, Stephen Shead, and Alan Fekete. A teaching system for SQL. In *Proceedings of the 2nd Australasian conference on Computer science education*, pages 224–231. ACM, 1997.

[Keim *et al.* 2008] Daniel A Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual analytics: Scope and challenges. In *Visual data mining*, pages 76–90. Springer, 2008.

[Keim *et al.* 2010] Daniel A Keim, Ming C Hao, Umeshwar Dayal, Halldor Janetzko, and Peter Bak. Generalized scatter plots. *Information Visualization*, 9(4):301–311, 2010.

[Kejriwal *et al.* 2017] Mayank Kejriwal, Jiayuan Ding, Runqi Shao, Anoop Kumar, and Pedro Szekely. Flagit: A system for minimally supervised human trafficking indicator mining. *arXiv preprint arXiv:1712.03086*, 2017.

[Keles *et al.* 2020] Betul Keles, Niall McCrae, and Annmarie Grealish. A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International journal of adolescence and youth*, 25(1):79–93, 2020.

[Kelih *et al.* 2005] Emmerich Kelih, Gordana Antić, Peter Grzybek, and Ernst Stadlober. Classification of author and/or genre? the impact of word length. In *Classificationâ€"The ubiquitous challenge*, pages 498–505. Springer, 2005.

[Kelleher 2013] David Kelleher. Survey: 81% of US Employees Check Their Work Mail Outside Work Hours. *TechTalk. Last Modified May*, 20, 2013.

[Kellems *et al.* 2016] Ryan O Kellems, Terisa P Gabrielsen, and Caroline Williams. Using visual organizers and technology: Supporting executive function, abstract language comprehension, and social learning. In *Technology and the Treatment of Children with Autism Spectrum Disorder*, pages 75–86. Springer, 2016.

[Kendall and Gal 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[Kennan *et al.* 2005] Mary Anne Kennan, Fletcher Cole, Patricia Willard, and C.S Wilson. Changes in the workplace: transformation of the information professional? In *2nd Research Applications in Information and Library Studies Seminar, Canberra*, pages 1–19, 2005.

[Keogh 2006] Eamonn Keogh. Naive Bayes classifier. 2006.

[Këpuska and Bohouta 2017] Veton Këpuska and Gamal Bohouta. Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *Int. J. Eng. Res. Appl*, 7(03):20–24, 2017.

[Keselj and Wang] Fuchun Pengt Dale Schuurmanst Vlado Keselj and Shaojun Wang. Language independent authorship attribution using character level language models.

[Keselj and Wang 2003] Fuchun Pengt Dale Schuurmanst Vlado Keselj and Shaojun Wang. Language independent authorship attribution using character level language models. In *EACL 2003: 10th Conference of the European Chapter, Association for Computational Linguistics-Proceedings of the Conference, April 12th-17th 2003, Agro Hotel, Budapest, Hungary*, page 267. Citeseer, 2003.

[Kešelj *et al.* 2003] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264, 2003.

[Keselj 2009] Vlado Keselj. *Speech and Language Processing Pearson Prentice Hall*, 2009.

[Kessler *et al.* 1997] Brett Kessler, Geoffrey Nunberg, and Hinrich Schuetze. Automatic detection of text genre. 1997.

[Kessler 2017] Jason S. Kessler. *Scattertext: a Browser-Based Tool for Visualizing how Corpora Differ*, 2017.

[Kestemont *et al.* 2018] Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. Overview of the author identification task at pan-2018: cross-domain authorship attribution and style change detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.*, pages 1–25, 2018.

[Kestemont *et al.* 2019] Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. Overview of the cross-domain authorship attribution task at pan 2019. In *CLEF*, 2019.

[Kestemont 2014] Mike Kestemont. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, 2014.

[Keyrouz *et al.* 2020] Youssef Keyrouz, Cyril Fonlupt, Dany Mezher, Denis Robilliard, and Rafic Faddoul. Evolving a weighted combination of text similarities for authorship attribution. In *International Conference on Artificial Evolution (Evolution Artificielle)*, pages 13–27. Springer, 2020.

[Khalid and Srinivasan 2020] Osama Khalid and Padmini Srinivasan. Style matters! investigating linguistic style in online communities. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 360–369, 2020.

[Khatun *et al.* 2019] Aisha Khatun, Anisur Rahman, Md. Saiful Islam, and Marium-E-Jannat. Authorship attribution in bangla literature using character-level cnn. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5, 2019.

[Khmelev and Teahan 2003] Dmitry V Khmelev and William J Teahan. A repetition based measure for verification of text collections and for text categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 104–110, 2003.

[Khomytska *et al.*] Iryna Khomytska, Vasyl Teslyuk, Andriy Holovatyy, and Oleksandr Morushko. Development of methods, models, and means for the author attribution of a text. *Eastern-European Journal of Enterprise Technologies*, 3(2):41–46.

[Khomytska *et al.* 2020] Iryna Khomytska, Vasyl Teslyuk, Iryna Bazylevych, and Inna Shylinska. Approach for minimization of phoneme groups in authorship attribution. *International Journal of Computing*, 19(1):55–62, 2020.

[Khondaker *et al.* 2020] Md Tawkat Islam Khondaker, Junaed Younus Khan, Tanvir Alam, and M Sohel Rahman. Agree-to-disagree (a2d): A deep learning-based framework for authorship discrimination task in corpus-specificity free manner. *IEEE Access*, 8:162322–162334, 2020.

[Khonji and Iraqi 2014] Mahmoud Khonji and Youssef Iraqi. A slightly-modified gi-based author-verifier with lots of features (asgalf). *CLEF (Working Notes)*, 1180:977–983, 2014.

[Khosravi *et al.* 2018] Pegah Khosravi, Ehsan Kazemi, Marcin Imielinski, Olivier Elemento, and Iman Hajirasouliha. Deep convolutional neural networks enable discrimination of heterogeneous digital pathology images. *EBioMedicine*, 27:317–328, 2018.

[Khurana *et al.* 2017] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *arXiv preprint arXiv:1708.05148*, 2017.

[Khuziakhmetov and Porchesku 2016] Anvar N Khuziakhmetov and Galina V Porchesku. Teaching listening comprehension: Bottom-up approach. *International journal of environmental and science education*, 11(8):1989–2001, 2016.

[Kietzmann *et al.* 2011] Jan H. Kietzmann, Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, 54(3):241–251, 2011. SPECIAL ISSUE: SOCIAL MEDIA.

[Kim *et al.* 2016] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *AAAI*, pages 2741–2749, 2016.

[Kim *et al.* 2020] Yoonsang Kim, Rachel Nordgren, and Sherry Emery. The story of goldilocks and three twitterâ€™s apis: a pilot study on twitter data sources and disclosure. *International Journal of Environmental Research and Public Health*, 17(3):864, 2020.

[Kim 2014] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Kinchin 2011] Ian M Kinchin. Visualising knowledge structures in biology: Discipline, curriculum and student understanding. *Journal of Biological Education*, 45(4):183–189, 2011.

[Kingma and Ba 2014] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. https://arxiv.org/abs/1412.6980, 2014.

[Kirillov 2017] Danil Kirillov. *Clustering Spam Emails Using Hadoop and FP-Trees*. PhD thesis, Carleton University, 2017.

[Kjell 1994] B. Kjell. Authorship attribution of text samples using neural networks and bayesian classifiers. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*, volume 2, pages 1660–1664 vol.2, Oct 1994.

[Klarreich 2003] Erica Klarreich. Statistical tests are unraveling knotty literary mysteries. *Science News*, 164(25):392, 2003.

[Kleerekoper and Schofield 2018] Anthony Kleerekoper and Andrew Schofield. SQL tester: an online SQL assessment tool and its impact. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 87–92. ACM, 2018.

[Klein *et al.* 2017] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, 2017.

[Klein 2006] Ewan Klein. Computational semantics in the natural language toolkit. In *Proceedings of the Australasian Language Technology Workshop 2006*, pages 26–33, 2006.

[Kleiner *et al.* 2013] Carsten Kleiner, Christopher Tebbe, and Felix Heine. Automated grading and tutoring of SQL statements to improve student learning. In *Proceedings of the 13th Koli Calling International Conference on Computing Education Research*, pages 161–168. ACM, 2013.

[Koehn and Knowles 2017] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.

[Koehn *et al.* 2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180, 2007.

[Kokkalis *et al.* 2012] Andreas Kokkalis, Panagiotis Vagenas, Alexandros Zervakis, Alkis Simitsis, Georgia Koutrika, and Yannis Ioannidis. Logos: a system for translating queries into narratives. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 673–676. ACM, 2012.

[Koppel and Schler 2003] Moshe Koppel and Jonathan Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, pages 72–80, 2003.

[Koppel and Schler 2004] Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM, 2004.

[Koppel and Winter 2014] Moshe Koppel and Yaron Winter. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187, 2014.

[Koppel *et al.* 2002] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and linguistic computing*, 17(4):401–412, 2002.

[Koppel *et al.* 2006a] Moshe Koppel, Navot Akiva, and Ido Dagan. Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology*, 57(11):1519–1525, 2006.

[Koppel *et al.* 2006b] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660. ACM, 2006.

[Koppel *et al.* 2009] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26, 2009.

[Koppel *et al.* 2011] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011.

[Koppel *et al.* 2012] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Yaron Winter. The fundamental problem of authorship attribution. *English Studies*, 93(3):284–291, 2012.

[Kornacker *et al.* 2015] Marcel Kornacker, Alexander Behm, Victor Bittorf, Taras Bobrovytsky, Casey Ching, Alan Choi, Justin Erickson, Martin Grund, Daniel Hecht, Matthew Jacobs, et al. Impala: A modern, open-source SQL engine for Hadoop. In *7th Biennial Conference on Innovative Data Systems Research*, volume 1, page 9, 2015.

[Kostyuchenko *et al.* 2018] Evgeny Yurievich Kostyuchenko, Ivan Viktorovich, Botna Renko, and Alexander Alexandrovich Shelupanov. User identification by the free-text keystroke dynamics. In *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*, pages 1–4. IEEE, 2018.

[Kotov 2017] Mykhailo Kotov. NLP resources for a rare language morphological analyzer: Danish case. In *Computational linguistics and intelligent systems (COLINS 2017)*. National Technical University «KhPI», 2017.

[Kotzé *et al.* 2008] Paula Kotzé, Karen Renaud, and Judy van Biljon. Donâ€™t do this–pitfalls in using anti-patterns in teaching human–computer interaction principles. *Computers & Education*, 50(3):979–1008, 2008.

[Kou *et al.* 2020] Gang Kou, Pei Yang, Yi Peng, Feng Xiao, Yang Chen, and Fawaz E Alsaadi. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Applied Soft Computing*, 86:105836, 2020.

[Kouemou and Dymarski 2011] Guy Leonard Kouemou and Dr Przemyslaw Dymarski. History and theoretical basics of hidden Markov models. *Hidden Markov Models, Theory and Applications*, 1, 2011.

[Kourtis and Stamatatos 2011] Ioannis Kourtis and Efstathios Stamatatos. Author identification using semi-supervised learning. In *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands*, 2011.

[Kowsari *et al.* 2019] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

[Kowsher *et al.* 2021] Md Kowsher, Anik Tahabilder, Md Zahidul Islam Sanjid, Nusrat Jahan Prottasha, Md Shihab Uddin, Md Arman Hossain, and Md Abdul Kader Jilani. Lstm-ann & bilstm-ann: Hybrid deep learning models for enhanced classification accuracy. *Procedia Computer Science*, 193:131–140, 2021.

[Krackhardt *et al.* 1994] David Krackhardt, Jim Blythe, and Cathleen McGrath. An improved network drawing program. *Connections*, 17(2):53–55, 1994.

[Kramsch 2004] Claire Kramsch. Language, thought, and culture. *The handbook of applied linguistics*, pages 235–261, 2004.

[Kraska 2018] Tim Kraska. Northstar: an interactive data science system. *Proceedings of the VLDB Endowment*, 11(12):2150–2164, 2018.

[Krathwohl and Anderson 2001] David R Krathwohl and Lorin W Anderson. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2001.

[Krathwohl 2002] David R Krathwohl. A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218, 2002.

[Kreie and Ernst 2013] Jennifer Kreie and Bruce A Ernst. From database concepts to application: Use problem-based learning and Oracle development tools to facilitate learning. In *Proceedings of the Information Systems Educators Conference*, volume 2167, page 1435. Citeseer, 2013.

[Krishnapuram *et al.* 2005] Balaji Krishnapuram, Lawrence Carin, Mário AT Figueiredo, and Alexander J Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE transactions on pattern analysis and machine intelligence*, 27(6):957–968, 2005.

[Křivka and Meduna 2015] Zbyněk Křivka and Alexander Meduna. Jumping grammars. *International Journal of Foundations of Computer Science*, 26(06):709–731, 2015.

[Krusche *et al.* 2018] Stephan Krusche, Bruce Scharlau, Åsa Cajander, and Janet Hughes. 50 years of software engineering: challenges, results, and opportunities in its education. In *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 362–363. ACM, 2018.

[Kugler 2019] Logan Kugler. Being recognized everywhere. *Communications of the ACM*, 62(2):17–19, 2019.

[Kulkarni *et al.* 2021] Sagar Kulkarni, Sharvari Govilkar, and Dhiraj Amin. Analysis of plagiarism detection tools and methods. In *Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021)*, 2021.

[Kullar *et al.* 2020] Ravina Kullar, Debra A. Goff, Timothy P. Gauthier, and Tara C. Smith. To tweet or not to tweetâ€"a review of the viral power of twitter for infectious diseases. *Current Infectious Disease Reports*, 22:1–6, 6 2020.

[Kumar and Kanal 1988] Vipin Kumar and Laveen N Kanal. The CDP: A unifying formulation for heuristic search, dynamic programming, and branch-and-bound. In *Search in Artificial Intelligence*, pages 1–27. Springer, 1988.

[Kumaresan and Palanisamy 2017] T Kumaresan and C Palanisamy. E-mail spam classification using s-cuckoo search and support vector machine. *International Journal of Bio-Inspired Computation*, 9(3):142–156, 2017.

[Kupferman 2018] Orna Kupferman. Automata theory and model checking. In *Handbook of Model Checking*, pages 107–151. Springer, 2018.

[Kuznetsov *et al.* 2016] Mikhail P Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, and Vadim V Strijov. Methods for intrinsic plagiarism detection and author diarization. In *CLEF (Working notes)*, pages 912–919, 2016.

[Kwak *et al.* 2010] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 591â€"600, New York, NY, USA, 2010. Association for Computing Machinery.

[Kwon *et al.* 2019] Youngsung Kwon, Alexis Kwasinski, and Andres Kwasinski. Solar irradiance forecast using naïve Bayes classifier based on publicly available weather forecasting variables. *Energies*, 12(8):1529, 2019.

[Kyfonidis *et al.* 2017] Charalampos Kyfonidis, Nektarios Moumoutzis, and Stavros Christodoulakis. Block-C: A block-based programming teaching tool to facilitate introductory C programming courses. In *Global Engineering Education Conference (EDUCON), 2017 IEEE*, pages 570–579. IEEE, 2017.

[Labov 2011] William Labov. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. John Wiley & Sons, 2011.

[Lachman 1997] Sheldon J Lachman. Learning is a process: Toward an improved definition of learning. *The Journal of psychology*, 131(5):477–480, 1997.

[Lackey *et al.* 2014] Amanda E Lackey, Mariam Moshiri, Tarun Pandey, Chandana Lall, Neeraj Lalwani, and Puneet Bhargava. Productivity, part 1: getting things done, using e-mail, scanners, reference managers, note-taking applications, and text expanders. *Journal of the American College of Radiology*, 11(5):481–489, 2014.

[Lagutina *et al.* 2019] Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and P.G. Demidov. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195, Nov 2019.

[Lahtinen *et al.* 2005] Essi Lahtinen, Kirsti Ala-Mutka, and Hannu-Matti Järvinen. A study of the difficulties of novice programmers. In *ACM SIGSCE Bulletin*, volume 37, pages 14–18. ACM, 2005.

[Lahtinen 2007] Essi Lahtinen. A categorization of novice programmers: A cluster analysis study. In *PPIG*, volume 16, pages 32–41, 2007.

[Lai *et al.* 2008] Jennifer Lai, Clare-Marie Karat, and Nicole Yankelovich. Conversational speech interfaces and technologies. *The Human-Computer Interaction Handbook,*, pages 381–391, 2008.

[Lai *et al.* 2015] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[Lakoff 1979] Robin T Lakoff. Stylistic strategies within a grammar of style. *Annals of the New York Academy of Sciences*, 1979.

[Lambers and Veenman 2009] Maarten Lambers and Cor J Veenman. Forensic authorship attribution using compression distances to prototypes. In *International Workshop on Computational Forensics*, pages 13–24. Springer, 2009.

[Lane 2012] H Chad Lane. Cognitive models of learning. *Encyclopedia of the sciences of learning*, pages 608–610, 2012.

[Largeron *et al.* 2011] Christine Largeron, Christophe Moulin, and Mathias Géry. Entropy based feature selection for text categorization. In *Proceedings of the 2011 ACM symposium on applied computing*, pages 924–928, 2011.

[Larsson and Moe 2012] Anders Olof Larsson and Hallvard Moe. Studying political microblogging: Twitter users in the 2010 swedish election campaign. *New media & society*, 14(5):729–747, 2012.

[Lauriat 2018] Barbara Lauriat. Literary and dramatic disputes in shakespeareâ€™s time. *Journal of International Dispute Settlement*, 9(1):45–57, 2018.

[Laurillard 2013] Diana Laurillard. *Teaching as a design science: Building pedagogical patterns for learning and technology*. Routledge, 2013.

[Lauw *et al.* 2010] Hady Lauw, John C Shafer, Rakesh Agrawal, and Alexandros Ntoulas. Homophily in the digital world: A livejournal case study. *IEEE Internet Computing*, 14(2):15–23, 2010.

[Lavbič *et al.* 2017] Dejan Lavbič, Tadej Matek, and Aljaž Zrnec. Recommender system for learning SQL using hints. *Interactive Learning Environments*, 25(8):1048–1064, 2017.

[Lawrence 2014] Ramon Lawrence. Integration and virtualization of relational SQL and NoSQL systems including mySQL and MongoDB. In *2014 International Conference on Computational Science and Computational Intelligence*, volume 1, pages 285–290. IEEE, 2014.

[Layton *et al.* 2013] Robert Layton, Paul Watters, and Richard Dazeley. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, 19(1):95–120, 2013.

[Layton *et al.* 2015] Robert Layton, Paul A Watters, and Richard Dazeley. Authorship analysis of aliases: Does topic influence accuracy? *Natural Language Engineering*, 21(4):497–518, 2015.

[LeCun *et al.* 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[Lee *et al.* 2013] Yun Young Lee, Nicholas Chen, and Ralph E Johnson. Drag-and-drop refactoring: intuitive and efficient program transformation. In *Proceedings of the 2013 International Conference on Software Engineering*, pages 23–32. IEEE Press, 2013.

[Lee *et al.* 2017] Jongchan Lee, Jae-Woong Choe, and Mingzhe Jin. Authorship attribution of korean texts by using phrase patterns. *International Information Institute (Tokyo). Information*, 20(1B):417, 2017.

[Leech and Smith 2005] Geoffrey Leech and Nicholas Smith. Extending the possibilities of corpus-based research on english in the twentieth century: A prequel to lob and flob. *ICAME Journal*, 29:83–98, 2005.

[Leech 1992] Geoffrey Neil Leech. 100 million words of English: the British National Corpus (BNC). 1992.

[Leeflang *et al.* 2014] Peter S.H. Leeflang, Peter C. Verhoef, Peter DahlstrÃ¶m, and Tjark Freundt. Challenges and solutions for marketing in a digital era. *European Management Journal*, 32(1):1–12, 2014.

[Lehnert and Ringle 2014] Wendy G Lehnert and Martin H Ringle. *Strategies for natural language processing*. Psychology Press, 2014.

[Lemay *et al.* 2019] David John Lemay, Ram B Basnet, Tenzin Doleck, and Paul Bazelais. Social network analysis of twitter use during the aera 2017 annual conference. *Education and Information Technologies*, 24:459–470, 2019.

[Lemut *et al.* 2013] Enrica Lemut, Benedict DuBoulay, and Giuliana Dettori. *Cognitive models and intelligent environments for learning programming*, volume 111. Springer Science & Business Media, 2013.

[Letovsky 1987] Stanley Letovsky. Cognitive processes in program comprehension. *Journal of Systems and software*, 7(4):325–339, 1987.

[Levene and Loizou 2012] Mark Levene and George Loizou. *A guided tour of relational databases and beyond*. Springer Science & Business Media, 2012.

[Levulis *et al.* 2018] Samuel J Levulis, Patricia R DeLucia, and So Young Kim. Effects of touch, voice, and multimodal input, and task load on multiple - UAV monitoring performance during simulated manned-unmanned teaming in a military helicopter. *Human factors*, 60(8):1117–1129, 2018.

[Lewis *et al.* 2004] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[Li and Jagadish 2014a] Fei Li and Hosagrahar V Jagadish. NaLIR: an interactive natural language interface for querying relational databases. In *Proceedings of the ACM SIGMOD international conference on Management of data*, pages 709–712. ACM, 2014.

[Li and Jagadish 2014b] Fei Li and HV Jagadish. Constructing an interactive natural language interface for relational databases. *Proceedings of the VLDB Endowment*, 8(1):73–84, 2014.

[Li and Jagadish 2016] Fei Li and HV Jagadish. Understanding natural language queries over relational databases. *ACM Special Interest Group on Management of Data Record*, 45(1):6–13, 2016.

[Li *et al.* 2008] Yunyao Li, Rajasekar Krishnamurthy, Sriram Raghavan, Shivakumar Vaithyanathan, and HV Jagadish. Regular expression learning for information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 21–30. Association for Computational Linguistics, 2008.

[Li *et al.* 2010] Linlin Li, Benjamin Roth, and Caroline Sporleder. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147. Association for Computational Linguistics, 2010.

[Li *et al.* 2016] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[Li *et al.* 2018] Tong Li, Jin Li, Zheli Liu, Ping Li, and Chunfu Jia. Differentially private naive Bayes learning over multiple data sources. *Information Sciences*, 444:89–104, 2018.

[Li *et al.* 2019] Jingjing Li, Wenlu Wang, Wei-Shinn Ku, Yingtao Tian, and Haixun Wang. SpatialNLI: A spatial domain natural language interface to databases using spatial comprehension. *arXiv preprint arXiv:1908.10917*, 2019.

[Li *et al.* 2020] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S Yu, and Lifang He. A survey on text classification: From shallow to deep learning. *arXiv preprint arXiv:2008.00364*, 2020.

[Liao *et al.* 2021] Shu-Hsien Liao, Retno Widowati, and Yu-Chieh Hsieh. Investigating online social media usersâ€™ behaviors for social commerce recommendations. *Technology in Society*, 66:101655, 2021.

[Liberty 2005] Jesse Liberty. *Programming C#: Building. NET Applications with C*. " O'Reilly Media, Inc.", 2005.

[Liddy 2001] Elizabeth D Liddy. Natural language processing. 2001.

[Lieberman *et al.* 2006] Henry Lieberman, Fabio Paternò, Markus Klann, and Volker Wulf. End-user development: An emerging paradigm. In *End user development*, pages 1–8. Springer, 2006.

[Lim *et al.* 2016] Wei Ying Lim, Angela Ong, Lay Lian Soh, and Adam Sufi. Teachersâ€™ voices and change: The structure and agency dialectics that shaped teachersâ€™ pedagogy toward deep learning. In *Future learning in primary schools*, pages 147–158. Springer, 2016.

[Lin *et al.* 2014] Yuming Lin, Tao Zhu, Hao Wu, Jingwei Zhang, Xiaoling Wang, and Aoying Zhou. Towards online anti-opinion spam: Spotting fake reviews from the review sequence. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 261–264. IEEE, 2014.

[Lin 2021] Ying Lin. *10 Blogging Statistics You Need to Know in 2021 [Infographic]*, 2021.

[Linz 2011] Peter Linz. *An introduction to formal languages and automata*. Jones & Bartlett Publishers, 2011.

[Lison and Kennington 2016] Pierre Lison and Casey Kennington. Opendial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *Proceedings of ACL-2016 System Demonstrations*, pages 67–72, 2016.

[Littman *et al.* 1986] David C Littman, Jeannine Pinto, Stanley Letovsky, and Elliot Soloway. *Mental models and software maintenance, Papers presented at the first workshop on empirical studies of programmers on Empirical studies of programmers (Norwood, NJ, USA)*, 1986.

[Litvinova *et al.* 2015] TA Litvinova, PV Seredin, and OA Litvinova. Using part-of-speech sequences frequencies in a text to predict author personality: a corpus study. *Indian Journal of Science and Technology*, 8:93, 2015.

[Liu and Guo 2019] Gang Liu and Jiabao Guo. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338, 2019.

[Liu and Ruths 2013] Wendy Liu and Derek Ruths. Whatâ€™s in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*, 2013.

[Liu and Yu 2005] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502, 2005.

[Liu *et al.* 2003] Xia Liu, Lai C Liu, Kai S Koong, and June Lu. An examination of job skills posted on Internet databases: Implications for information systems degree programs. *Journal of Education for Business*, 78(4):191–196, 2003.

[Liu *et al.* 2011] Chen-Chung Liu, Yuan-Bang Cheng, and Chia-Wen Huang. The effect of simulation games on the learning of computational problem solving. *Computers & Education*, 57(3):1907–1918, 2011.

[Liu *et al.* 2017] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, 2017.

[Liu *et al.* 2019] Qian Liu, Bei Chen, Jian-Guang Lou, Ge Jin, and Dongmei Zhang. Fanda: A novel approach to perform follow-up query analysis. *arXiv preprint arXiv:1901.08259*, 2019.

[Liu *et al.* 2020] Bing Liu, Yong Zhou, and Wei Sun. Character-level text classification via convolutional neural network and gated recurrent unit. *International Journal of Machine Learning and Cybernetics*, 11(8):1939–1949, 2020.

[Lobur *et al.* 2011] Mykhailo Lobur, AnDriy Romanyuk, and Mariana Romanyshyn. Using nltk for educational and scientific purposes. In *2011 11th international conference the experience of designing and application of CAD systems in microelectronics (CADSM)*, pages 426–428. IEEE, 2011.

[Loper and Bird 2002] Edward Loper and Steven Bird. NLTK: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[López *et al.* 2017] Gustavo López, Luis Quesada, and Luis A Guerrero. Alexa vs. Siri vs. Cortana vs. Google assistant: A comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*, pages 241–250. Springer, 2017.

[Lopez-Gazpio *et al.* 2019] Inigo Lopez-Gazpio, Montse Maritxalar, Mirella Lapata, and Eneko Agirre. Word n-gram attention models for sentence similarity and inference. *Expert Systems with Applications*, 132:1–11, 2019.

[López-Monroy *et al.* 2020] A Pastor López-Monroy, Fabio A Gonzalez, and Thamar Solorio. Early author profiling on twitter using profile features with multi-resolution. *Expert Systems with Applications*, 140:112909, 2020.

[Lorenz *et al.* 2019] Klara Lorenz, Paul P Freddolino, Adelina Comas-Herrera, Martin Knapp, and Jacqueline Damant. Technology-based tools and services for people with dementia and carers: Mapping technology onto the dementia care pathway. *Dementia*, 18(2):725–741, 2019.

[Loshchilov and Hutter 2017] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*, 2017.

[Lowe *et al.* 2016] Matthew X Lowe, Ryan A Stevenson, Kristin E Wilson, Natasha E Ouslis, Morgan D Barense, Jonathan S Cant, and Susanne Ferber. Sensory processing patterns predict the integration of information held in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 42(2):294, 2016.

[Lu *et al.* 2012] Meiyu Lu, Srinivas Bangalore, Graham Cormode, Marios Hadjieleftheriou, and Divesh Srivastava. A dataset search engine for the research document corpus. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1237–1240. IEEE, 2012.

[Lu *et al.* 2015] Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, and Guangquan Zhang. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems*, 80:14–23, 2015.

[Lundberg and Lee 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[Luque *et al.* 2019] Amalia Luque, Alejandro Carrasco, Alejandro Martín, and Ana de las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.

[Luxton-Reilly *et al.* 2018] Andrew Luxton-Reilly, Ibrahim Albluwi, Brett A Becker, Michail Giannakos, Amruth N Kumar, Linda Ott, James Paterson, Michael James Scott, Judy Sheard, Claudia Szabo, et al. Introductory programming: a systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, pages 55–106. ACM, 2018.

[Luyckx and Daelemans 2008] Kim Luyckx and Walter Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 513–520. Association for Computational Linguistics, 2008.

[Luyckx and Daelemans 2011] Kim Luyckx and Walter Daelemans. The effect of author set size and data size in authorship attribution. *Literary and linguistic Computing*, 26(1):35–55, 2011.

[Lye and Koh 2014] Sze Yee Lye and Joyce Hwee Ling Koh. Review on teaching and learning of computational thinking through programming: What is next for k-12? *Computers in Human Behavior*, 41:51–61, 2014.

[Lyons *et al.* 2016] Gabriel Lyons, Vinh Tran, Carsten Binnig, Ugur Cetintemel, and Tim Kraska. Making the case for Query by Voice with Echoquery. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2129–2132. ACM, 2016.

[M. Otala *et al.* 2021] Jacqueline M. Otala, Gillian Kurtic, Isabella Grasso, Yu Liu, Jeanna Matthews, and Golshan Madraki. Political polarization and platform migration: A study of parler and twitter usage by united states of america congress members. In *Companion Proceedings of the Web Conference 2021*, pages 224–231, 2021.

[Ma *et al.* 2020] Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. Towards improved model design for authorship identification: A survey on writing style understanding. *CoRR*, abs/2009.14445, 2020.

[MacLeod and Grant 2012] Nicci MacLeod and Tim Grant. Whose tweet? authorship analysis of micro-blogs and other short-form messages. 2012.

[Madigan *et al.* 2005a] David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. Author identification on the large scale. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*, 2005.

[Madigan *et al.* 2005b] David Madigan, Alexander Genkin, David D Lewis, and Dmitriy Fradkin. Bayesian multinomial logistic regression for author identification. In *AIP conference proceedings*, volume 803, pages 509–516. American Institute of Physics, 2005.

[Madnani 2007] Nitin Madnani. Getting started on natural language processing with Python. *ACM Crossroads*, 13(4):5, 2007.

[Maguluri *et al.* 2019] Lakshmana Phaneendra Maguluri, R Ragupathy, Sita Rama Krishna Buddi, Vamshi Ponugoti, and Tharun Sai Kalimil. Adaptive prediction of spam emails: Using Bayesian inference. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 628–632. IEEE, 2019.

[Mahmood *et al.* 2020] Asad Mahmood, Zubair Shafiq, and Padmini Srinivasan. A girl has a name: Detecting authorship obfuscation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2235–2245, 2020.

[Maier and Größler 2000] Frank H Maier and Andreas Größler. What are we talking about?â€"a taxonomy of computer simulations to support learning. *System Dynamics Review: The Journal of the System Dynamics Society*, 16(2):135–148, 2000.

[Majeed and Lee 2020] Abdul Majeed and Sungchang Lee. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE access*, 9:8512–8545, 2020.

[Makridakis 2017] Spyros Makridakis. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, 90:46–60, 2017.

[Malan and Leitner 2007] David J Malan and Henry H Leitner. Scratch for budding computer scientists. In *ACM Special Interest Group on Computer Science Education Bulletin*, volume 39, pages 223–227. ACM, 2007.

[Malhotra and Malhotra 2016] Claudia Kubowicz Malhotra and Arvind Malhotra. How ceos can leverage twitter. *MIT Sloan Management Review*, 57(2):73, 2016.

[Malyutov 2005] Mikhail B Malyutov. Authorship attribution of texts: a review. *Electronic Notes in Discrete Mathematics*, 21:353–357, 2005.

[Manca 2020] Stefania Manca. Snapping, pinning, liking or texting: Investigating social media in higher education beyond facebook. *The Internet and Higher Education*, 44:100707, 2020.

[Mannila *et al.* 2014] Linda Mannila, Valentina Dagiene, Barbara Demo, Natasa Grgurina, Claudio Mirolo, Lennart Rolandsson, and Amber Settle. Computational thinking in k-9 education. In *Proceedings of the working group reports of the 2014 on innovation & technology in computer science education conference*, pages 1–29. ACM, 2014.

[Manning *et al.* 1999] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[Manning *et al.* 2010] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[Manning *et al.* 2014] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[Manning 2014] Christopher Manning. Natural language processing. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. Citeseer, 2014.

[Manning 2015] Christopher D Manning. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707, 2015.

[Mannino 2001] Michael Mannino. *Database application development and design*. McGraw-Hill, Inc., 2001.

[Mao and Brown 2007] Ji-Ye Mao and Bradley R Brown. The effectiveness of online task support vs. instructor-led training. *Contemporary Issues in End User Computing*, page 77, 2007.

[Mao *et al.* 2018] Rui Mao, Chenghua Lin, and Frank Guerin. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL), 2018.

[Mao *et al.* 2023] Yuanyi Mao, Tianyi Xu, and Ki Joon Kim. Motivations for proactive and reactive trolling on social media: Developing and validating a four-factor model. *Social Media+ Society*, 9(4):20563051231203682, 2023.

[Marcu 2000] Daniel Marcu. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational linguistics*, 26(3):395–448, 2000.

[Marcus *et al.* 1994] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics, 1994.

[Marcus *et al.* 2019] Ryan Marcus, Parimarjan Negi, Hongzi Mao, Chi Zhang, Mohammad Alizadeh, Tim Kraska, Olga Papaemmanouil, and Nesime Tatbul. Neo: A learned query optimizer. *arXiv preprint arXiv:1904.03711*, 2019.

[Marginean 2017] Anca Marginean. Question answering over biomedical linked data with grammatical framework. *Semantic Web*, 8(4):565–580, 2017.

[Markov *et al.* 2017] Ilia Markov, Efstathios Stamatatos, and Grigori Sidorov. Improving cross-topic authorship attribution: The role of pre-processing. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 289–302. Springer, 2017.

[Marom *et al.* 2010] Nadav David Marom, Lior Rokach, and Armin Shmilovici. Using the confusion matrix for improving ensemble classifiers. In *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pages 000555–000559, 2010.

[Martin 1991] John C Martin. *Introduction to Languages and the Theory of Computation*, volume 4. McGraw-Hill NY, 1991.

[Martin 2003] J.C. Martin. *Introduction to Languages and the Theory of Computation*. McGraw-Hill, New York, 2003.

[Martins *et al.* 2014] Carolina Martins, Tiago Oliveira, and Aleš Popovič. Understanding the Internet banking adoption: A unified theory of acceptance and use of technology and perceived risk application. *International Journal of Information Management*, 34(1):1–13, 2014.

[Marton *et al.* 2005] Yuval Marton, Ning Wu, and Lisa Hellerstein. On compression-based text classification. In *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*, pages 300–314. Springer, 2005.

[Mason *et al.* 2016] Raina Mason, Carolyn Seton, and Graham Cooper. Applying cognitive load theory to the redesign of a conventional database systems course. *Computer Science Education*, 26(1):68–87, 2016.

[Mathur *et al.* 2020] Amrita Mathur, Purnima Kubde, and Sonali Vaidya. Emotional analysis using twitter data during pandemic situation: Covid-19. In *2020 5th international conference on communication and electronics systems (ICCES)*, pages 845–848. IEEE, 2020.

[Maubert and Pinchinat 2013] Bastien Maubert and Sophie Pinchinat. Jumping automata for uniform strategies. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2013)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

[Maurer *et al.* 2006] Hermann Maurer, Frank Kappe, and Bilal Zaka. Plagiarism–a survey. *Journal of Universal Computer Science*, 12(8):1050–1084, 2006.

[Mayer and Alexander 2016] Richard E Mayer and Patricia A Alexander. *Handbook of research on learning and instruction*. Taylor & Francis, 2016.

[Mayer 2009] Richard E Mayer. Constructivism as a theory of learning versus constructivism as a prescription for instruction. *Constructivist instruction: Success or failure*, pages 184–200, 2009.

[Maynor 1994] Natalie Maynor. 7. the language of electronic mail: Written speech? *Publication of the American Dialect Society*, 78(1):48–54, 1994.

[Mayor *et al.* 2014] Cristhian Mayor, Josué Gerardo Gutiérrez Hernández, Angel Israel Toledo Castro, Rodrigo Martinez, Paola Ledesma, Gibran Fuentes, and Iván Vladimir Meza Ruíz. A single author style representation for the author verification task. In *CLEF (Working Notes)*, pages 1079–1083, 2014.

[Mayoraz and Alpaydin 1999] Eddy Mayoraz and Ethem Alpaydin. Support vector machines for multi-class classification. In *International Work-Conference on Artificial Neural Networks*, pages 833–842. Springer, 1999.

[McCann *et al.* 2016] James McCann, Lea Albaugh, Vidya Narayanan, April Grow, Wojciech Matusik, Jennifer Mankoff, and Jessica Hodgins. A compiler for 3D machine knitting. *ACM Transactions on Graphics*, 35(4):49, 2016.

[McDougle *et al.* 2016] Samuel D McDougle, Richard B Ivry, and Jordan A Taylor. Taking aim at the cognitive side of learning in sensorimotor adaptation tasks. *Trends in cognitive sciences*, 20(7):535–544, 2016.

[McGill 2008] Monica McGill. Critical skills for game developers: an analysis of skills sought by industry. In *Proceedings of the 2008 conference on future play: Research, play, share*, pages 89–96. ACM, 2008.

[McLeod 2003] Gregory McLeod. Learning theory and instructional design. *Learning Matters*, 2(3):35–43, 2003.

[McMenamin 2001] Gerald R McMenamin. Style markers in authorship studies. *International Journal of Speech, Language and the Law*, 8(2):93–97, 2001.

[McMenamin 2002] Gerald R McMenamin. *Forensic linguistics: Advances in forensic stylistics*. CRC press, 2002.

[McTear 2002] Michael F McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, 34(1):90–169, 2002.

[Mealin and Murphy-Hill 2012] Sean Mealin and Emerson Murphy-Hill. An exploratory study of blind software developers. In *2012 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 71–74. IEEE, 2012.

[Meduna and Soukup 2017] Alexander Meduna and Ondřej Soukup. *Modern Language Models and Computation: Theory with Applications*. Springer, 2017.

[Meduna and Zemek 2012] Alexander Meduna and Petr Zemek. Jumping finite automata. *International Journal of Foundations of Computer Science*, 23(07):1555–1578, 2012.

[Meduna and Zemek 2014] Alexander Meduna and Petr Zemek. *Regulated Grammars and Automata*. Springer, 2014.

[Meduna 2014] Alexander Meduna. *Formal Languages and Computation: Models and their applications*. Auerbach Publications, 2014.

[Mehmood *et al.* 2017] Muhammad Amir Mehmood, Hafiz Muhammad Shafiq, and Abdul Waheed. Understanding regional context of world wide web using common crawl corpus. In *2017 IEEE 13th Malaysia international conference on communications (MICC)*, pages 164–169. IEEE, 2017.

[Mekala *et al.* 2018] Sreenivas Mekala, Vishnu Vardan Bulusu, and Raghunadha Reddy. A survey on authorship attribution approaches. *Int. J. Comput. Eng. Res.(IJCER)*, 8(8), 2018.

[Melton 1998] Jim Melton. Database language SQL. In *Handbook on Architectures of Information Systems*, pages 103–128. Springer, 1998.

[Mendenhall 1887a] T. C. Mendenhall. The characteristic curves of composition. *Science*, ns-9(214s):237–246, 1887.

[Mendenhall 1887b] Thomas Corwin Mendenhall. The characteristic curves of composition. *Science*, (214s):237–246, 1887.

[Menekse 2019] Muhsin Menekse. The reflection-informed learning and instruction to improve studentsâ€™ academic success in undergraduate classrooms. *The Journal of Experimental Education*, pages 1–17, 2019.

[Merriam *et al.* 2006] Sharan B Merriam, Rosemary S Caffarella, and Lisa M Baumgartner. *Learning in adulthood: A comprehensive guide*. John Wiley & Sons, 2006.

[Metcalfe and Xu 2017] Janet Metcalfe and Judy Xu. Learning from oneâ€™s own errors and those of others. *Psychonomic Bulletin & Review*, pages 1–7, 2017.

[Meziane *et al.* 2008] Farid Meziane, Nikos Athanasakis, and Sophia Ananiadou. Generating natural language specifications from uml class diagrams. *Requirements Engineering*, 13(1):1–18, 2008.

[Michell 2013] Colin Simon Michell. *Investigating the use of forensic stylistic and stylometric techniques in the analyses of authorship on a publicly accessible social networking site (facebook)*. Master dissertation, Haverford College, 2013.

[Mihailidis 2014] Paul Mihailidis. The civic-social media disconnect: exploring perceptions of social media for engagement in the daily life of college students. *Information, Communication & Society*, 17(9):1059–1071, 2014.

[Mihalcea and Tarau 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[Miklody *et al.* 2017] Daniel Miklody, Wendie M Uitterhoeve, Dimitri van Heel, Kerstin Klinkenberg, and Benjamin Blankertz. Maritime cognitive workload assessment. *10.1007/978-3-319-57753-1*, 2017.

[Mikolov *et al.* 2013a] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[Mikolov *et al.* 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Mikros and Perifanos 2013] George K Mikros and Kostas Perifanos. Authorship attribution in greek tweets using author's multilevel n-gram profiles. In *2013 AAAI Spring Symposium Series*, 2013.

[Miller *et al.* 2017] Alexander H Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017.

[Miller 1995] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[Miller 2020] Daniel Miller. Social networking sites. In *Digital anthropology*, pages 146–161. Routledge, 2020.

[Milo and Somech 2018] Tova Milo and Amit Somech. Next-step suggestions for modern interactive data analysis platforms. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 576–585, 2018.

[Milo and Somech 2020] Tova Milo and Amit Somech. Automating exploratory data analysis via machine learning: An overview. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 2617–2622, 2020.

[Minnen *et al.* 2021] Molly E Minnen, Tanya Mitropoulos, Alexa K Rosenblatt, and Charles Calderwood. The incessant inbox: Evaluating the relevance of after-hours e-mail characteristics for work-related rumination and well-being. *Stress and Health*, 37(2):341–352, 2021.

[Mira 2008] José Mira Mira. Symbols versus connections: 50 years of artificial intelligence. *Neurocomputing*, 71(4-6):671–680, 2008.

[Mishne 2007] GA Mishne. *Applied Text Analytics for Blogs, Universiteit van Amsterdam*. PhD thesis, Dissertation, 27 April 2007 (Zitiert auf Seite 49.), 2007.

[Mishra and Jain 2016] Amit Mishra and Sanjay Kumar Jain. A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, 28(3):345–361, 2016.

[Mitra *et al.* 2007] Vikramjit Mitra, Chia-Jiu Wang, and Satarupa Banerjee. Text classification: A least square support vector machine approach. *Applied soft computing*, 7(3):908–914, 2007.

[Mitrovic and Ohlsson 2016] Antonija Mitrovic and Stellan Ohlsson. Implementing CBM: SQL-tutor after fifteen years. *International Journal of Artificial Intelligence in Education*, 26(1):150–159, 2016.

[Mitrovic 1998] Antonija Mitrovic. Learning SQL with a computerized tutor. In *ACM Special Interest Group on Computer Science Education Bulletin*, volume 30, pages 307–311. ACM, 1998.

[Mitrovic 2003] Antonija Mitrovic. An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education*, 13(2-4):173–197, 2003.

[Modaresi and Gross 2014] Pashutan Modaresi and Philipp Gross. A language independent author verifier using fuzzy c-means clustering. In *CLEF (Working Notes)*, pages 1084–1091, 2014.

[Modupe *et al.* 2011a] Abiodun Modupe, Oludayo O Olugbara, and Sunday O Ojo. Exploring support vector machines and random forests to detect advanced fee fraud activities on internet. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 331–335. IEEE, 2011.

[Modupe *et al.* 2011b] Abiodun Modupe, Oludayo O Olugbara, and Sunday O Ojo. Exploring support vector machines and random forests to detect advanced fee fraud activities on internet. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 331–335. IEEE, 2011.

[Modupe *et al.* 2012a] Abiodun Modupe, Oludayo O Olugbara, and Sunday O Ojo. Comparing Supervised Learning Classifiers to Detect Advanced Fee Fraud Activities on Internet. In *International Conference on Computer Science and Information Technology*, pages 87–100. Springer, Berlin, Heidelberg, 2012.

[Modupe *et al.* 2012b] Abiodun Modupe, Oludayo O Olugbara, and Sunday O Ojo. Comparing supervised learning classifiers to detect advanced fee fraud activities on internet. In *International Conference on Computer Science and Information Technology*, pages 87–100. Springer, 2012.

[Modupe *et al.* 2013a] Abiodun Modupe, Oludayo O Olugbara, and Sunday O Ojo. Investigating topic models for mobile short messaging service communication filtering. In *Proceedings of the world congress on engineering*, volume 2, pages 1–8, 2013.

[Modupe *et al.* 2013b] Abiodun Modupe, Oludayo O Olugbara, and Sunday O Ojo. Investigating topic models for mobile short messaging service communication filtering. In *Proceedings of the world congress on engineering*, volume 2, pages 1–8, 2013.

[Modupe *et al.* 2014] Abiodun Modupe, Oludayo O Olugbara, and Sunday O Ojo. Filtering of mobile short messaging service communication using latent dirichlet allocation with social network analysis. In *Transactions on Engineering Technologies*, pages 671–686. Springer, 2014.

[Modupe *et al.* 2022] Abiodun Modupe, Turgay Celik, Vukosi Marivate, and Oludayo O Olugbara. Post-authorship attribution using regularized deep neural network. *Applied Sciences*, 12(15):7518, 2022.

[Mogensen 2009] Torben Ægidius Mogensen. *Basics of Compiler Design*. Torben Ægidius Mogensen, 2009.

[Mohammad and Kiritchenko 2015] Saif M. Mohammad and Svetlana Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

[Mohammad and Turney 2010] Saif Mohammad and Peter Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34, 2010.

[Mohammed *et al.* 2017] Hussein Mohammed, Volker Märgner, Thomas Konidaris, and H Siegfried Stiehl. Normalised local naïve Bayes nearest-neighbour classifier for offline writer identification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1013–1018. IEEE, 2017.

[Mohan *et al.* 2010] Ashwin Mohan, Ibrahim M Baggili, and Marcus K Rogers. Authorship attribution of sms messages using an n-grams approach. *Proceedings of CERIAS Tech Report*, 11:1–12, 2010.

[Mohtashami and Scher 2000] Mojgan Mohtashami and Julian M Scher. Application of Bloomâ€™s cognitive domain taxonomy to database design. In *Proceedings of ISECON (information systems educators conference)*. Citeseer, 2000.

[Mohtasseb *et al.* 2009] Haytham Mohtasseb, Amr Ahmed, et al. Mining online diaries for blogger identification. 2009.

[Molnar 2020] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[Monroe *et al.* 2008] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008.

[Montazemi and Qahri-Saremi 2015] Ali Reza Montazemi and Hamed Qahri-Saremi. Factors affecting adoption of online banking: A meta-analytic structural equation modeling study. *Information & Management*, 52(2):210–226, 2015.

[Mony *et al.* 2014] Manju Mony, Jyothi M Rao, and Manish M Potey. An overview of NLIDB approaches and implementation for airline reservation system. *International Journal of Computer Applications*, 107(5), 2014.

[Moreno-León and Robles 2016] Jesús Moreno-León and Gregorio Robles. Code to learn with scratch? a systematic literature review. In *2016 IEEE Global Engineering Education Conference (EDUCON)*, pages 150–156. IEEE, 2016.

[Morgan and Sonquist 1963] James N Morgan and John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434, 1963.

[Morris 1970] Charles W Morris. The pragmatic movement in american philosophy. 1970.

[Morrison *et al.* 2013] Margaret A Morrison, Hyuk Jun Cheong, and Sally J McMillan. Posting, lurking, and networking: behaviors and characteristics of consumers in the context of user-generated content. *Journal of Interactive Advertising*, 13(2):97–108, 2013.

[Morton 1978] Andrew Queen Morton. Literary detection: How to prove authorship and fraud in literature and documents. *(No Title)*, 1978.

[Mössenböck 2005] Hanspeter Mössenböck. The Compiler Generator Coco/R User Manual. pages 1–33, 2005.

[Mosteller and Wallace 1963] Frederick Mosteller and David L Wallace. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[Mosteller and Wallace 1984] Frederick Mosteller and David L. Wallace. *The State of Statistical Authorship Studies in 1984*, pages 268–281. Springer New York, New York, NY, 1984.

[Mowrer 1960] Orval Mowrer. Learning theory and behavior. 1960.

[MSDN] MSDN. *Regular Expression Language - Quick Reference: Microsoft Developer Network Documentation on the .Net framework, version 4.5, regular expression.* https://msdn.microsoft.com/en-us/library/az24scfc(v=vs.110).aspx. Accessed: 2019-03-28.

[Muhammad *et al.* 2016] Aminu Muhammad, Nirmalie Wiratunga, and Robert Lothian. Contextual sentiment analysis for social media genres. *Knowledge-Based Systems*, 108:92–101, 2016.

[Muller 2019] Peter Muller. Style change detection. *ETH Zurich*, 2019.

[Mulliner *et al.* 2013] Collin Mulliner, Ravishankar Borgaonkar, Patrick Stewin, and Jean-Pierre Seifert. Sms-based one-time passwords: Attacks and defense: (short paper). In *Detection of Intrusions and Malware, and Vulnerability Assessment: 10th International Conference, DIMVA 2013, Berlin, Germany, July 18-19, 2013. Proceedings 10*, pages 150–159. Springer, 2013.

[Munteanu and Marcu 2005] Dragos Stefan Munteanu and Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 2005.

[Murauer and Specht 2021] Benjamin Murauer and Günther Specht. Developing a benchmark for reducing data bias in authorship attribution. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 179–188, 2021.

[Musiał and Kazienko 2013] Katarzyna Musiał and Przemysław Kazienko. Social networks on the internet. *World Wide Web*, 16:31–72, 2013.

[Muttenthaler *et al.* 2019] Lukas Muttenthaler, Gordon Lucas, and Janek Amann. Authorship attribution in fan-fictional texts given variable length character and word n-grams. In *CLEF (Working Notes)*, 2019.

[Myalapalli and Shiva 2015] Vamsi Krishna Myalapalli and Muddu Butchi Shiva. An appraisal to optimize SQL queries. In *International Conference on Pervasive Computing)*, pages 1–6. IEEE, 2015.

[Myers 1998] Brad A Myers. A brief history of human computer interaction technology. *interactions*, 5(2):44–54, 1998.

[Myung 2003] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100, 2003.

[Nadkarni *et al.* 2011] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[Naeem 2021] Muhammad Naeem. Do social media platforms develop consumer panic buying during the fear of covid-19 pandemic. *Journal of Retailing and Consumer Services*, 58:102226, 2021.

[Nagataki *et al.* 2013] Hiroyuki Nagataki, Yoshiaki Nakano, Midori Nobe, Tatsuya Tohyama, and Susumu Kanemune. A visual learning tool for database operation. In *Proceedings of the 8th Workshop in Primary and Secondary Computing Education*, pages 39–40. ACM, 2013.

[Nagy and Tick 2016] Albert Nagy and József Tick. Improving transport management with big data analytics. In *IEEE 14th International Symposium on Intelligent Systems and Informatics*, pages 199–204. IEEE, 2016.

[Naik *et al.* 2015] Ramesh R Naik, Maheshkumar B Landge, and C Namrata Mahender. A review on plagiarism detection tools. *International Journal of Computer Applications*, 125(11), 2015.

[Najar *et al.* 2014] Amir Shareghi Najar, Antonija Mitrovic, and Bruce M McLaren. Adaptive support versus alternating worked examples and tutored problems: Which leads to better learning? In *International Conference on User Modeling, Adaptation, and Personalization*, pages 171–182. Springer, 2014.

[Najar *et al.* 2016] Amir Shareghi Najar, Antonija Mitrovic, and Bruce M McLaren. Learning with intelligent tutors and worked examples: selecting learning activities adaptively leads to better learning outcomes than a fixed curriculum. *User Modeling and User-Adapted Interaction*, 26(5):459–491, 2016.

[Nakashima and Timberg 2017] Ellen Nakashima and Craig Timberg. Nsa officials worried about the day its potent hacking tool would get loose. then it did. *Washington Post*, 16, 2017.

[Nakazawa *et al.* 2006] Toshiaki Nakazawa, Kun Yu, Daisuke Kawahara, and Sadao Kurohashi. Example-based machine translation based on deeper NLP. In *International Workshop on Spoken Language Translation (IWSLT) 2006*, 2006.

[Náplava and Straka 2019] Jakub Náplava and Milan Straka. CUNI system for the building educational applications 2019 shared task: Grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 183–190, 2019.

[Narasimhan *et al.* 2016] Vagheesh Narasimhan, Petr Danecek, Aylwyn Scally, Yali Xue, Chris Tyler-Smith, and Richard Durbin. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32(11):1749–1751, 2016.

[Narayanan and Shmatikov 2009] Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy*, pages 173–187. IEEE, 2009.

[Nardi *et al.* 2000] Bonnie A. Nardi, Steve Whittaker, and Erin Bradner. Interaction and outeraction: Instant messaging in action. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 79–88, New York, NY, USA, 2000. ACM.

[Nardi *et al.* 2004a] Bonnie A Nardi, Diane J Schiano, Michelle Gumbrecht, and Luke Swartz. Why we blog. *Communications of the ACM*, 47(12):41–46, 2004.

[Nardi *et al.* 2004b] Bonnie A Nardi, Diane J Schiano, Michelle Gumbrecht, and Luke Swartz. Why we blog. *Communications of the ACM*, 47(12):41–46, 2004.

[Narula *et al.* 2008] Prayag Narula, Sanjay Kumar Dhurandher, Sudip Misra, and Isaac Woungang. Security in mobile ad-hoc networks using soft encryption and trust-based multipath routing. *Computer Communications*, 31(4):760–769, 2008.

[Naseem *et al.* 2021] Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(5), jun 2021.

[Nasukawa and Yi 2003] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM, 2003.

[Neal *et al.* 2017a] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, 50(6):1–36, 2017.

[Neal *et al.* 2017b] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36, 2017.

[Neal *et al.* 2018] Tempestt Neal, Kalaivani Sundararajan, and Damon Woodard. Exploiting linguistic style as a cognitive biometric for continuous verification. In *2018 International Conference on Biometrics (ICB)*, pages 270–276, 2018.

[Nelimarkka *et al.* 2021] Matti Nelimarkka, Teemu Leinonen, Eva Durall, and Philip Dean. Facebook is not a silver bullet for teachersâ€™ professional development: Anatomy of an eight-year-old social-media community. *Computers & Education*, 173:104269, 2021.

[Neteler *et al.* 2012] Markus Neteler, M Hamish Bowman, Martin Landa, and Markus Metz. GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software*, 31:124–130, 2012.

[Neumann and Kemper 2015] Thomas Neumann and Alfons Kemper. Unnesting arbitrary queries. *Datenbanksysteme für Business, Technologie und Web (BTW 2015)*, 2015.

[Newaz *et al.* 2021] Akm Iqtidar Newaz, Amit Kumar Sikder, Mohammad Ashiqur Rahman, and A Selcuk Uluagac. A survey on security and privacy issues in modern healthcare systems: Attacks and defenses. *ACM Transactions on Computing for Healthcare*, 2(3):1–44, 2021.

[Ng and Low 2004] Hwee Tou Ng and Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 277–284, 2004.

[Ng and Zelle 1997] Hwee Tou Ng and John Zelle. Corpus-based approaches to semantic interpretation in NLP. *AI magazine*, 18(4):45–45, 1997.

[Ng *et al.* 1997] Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 67–73, 1997.

[Nguyen *et al.* 2016] Nhu H Nguyen, Zewei Song, Scott T Bates, Sara Branco, Leho Tedersoo, Jon Menke, Jonathan S Schilling, and Peter G Kennedy. Funguild: an open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology*, 20:241–248, 2016.

[Nieuwland *et al.* 2007] Mante S Nieuwland, Marte Otten, and Jos JA Van Berkum. Who are you talking about? tracking discourse-level referential processing with event-related brain potentials. *Journal of Cognitive Neuroscience*, 19(2):228–236, 2007.

[Nigam and Ghani 2000] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 86–93, 2000.

[Nihalani *et al.* 2011] Neelu Nihalani, Sanjay Silakari, and Mahesh Motwani. Natural language interface for database: a brief review. *International Journal of Computer Science Issues (IJCSI)*, 8(2):600, 2011.

[Niu and Bansal 2018a] Tong Niu and Mohit Bansal. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389, 2018.

[Niu and Bansal 2018b] Tong Niu and Mohit Bansal. Polite Dialogue Generation Without Parallel Data. *Transactions of the Association for Computational Linguistics*, 6:373–389, 07 2018.

[Niu *et al.* 2017] Xing Niu, Marianna Martindale, and Marine Carpuat. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, 2017.

[Nordlinger and Sadler 2019] Rachel Nordlinger and Louisa Sadler. *Morphology in lexical-functional grammar and head-driven phrase structure grammar*. The Oxford handbook of morphological theory, 2019.

[Norouzifard *et al.* 2008] M Norouzifard, SH Davarpanah, MH Shenassa, et al. Using natural language processing in order to create SQL queries. In *International Conference on Computer and Communication Engineering*, pages 600–604. IEEE, 2008.

[Nzeyimana and Rubungo 2022] Antoine Nzeyimana and Andre Niyongabo Rubungo. Kinyabert: a morphology-aware kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, 2022.

[Oâ€™reilly 2005] Tim Oâ€™reilly. Design patterns and business models for the next generation of software. *Disponibile al sito: http://oreilly. com/web2/archive/what-is-web-20. html (21/02/2010)*, 2005.

[Ogden 1986] William C Ogden. Implications of a cognitive model of database query: comparison of a natural language, formal language and direct manipulation interface. *ACM SIGCHI Bulletin*, 18(2):51–54, 1986.

[Oh and Syn 2015] Sanghee Oh and Sue Yeon Syn. Motivations for sharing information and social support in social media: A comparative analysis of f acebook, t witter, d elicious, y ou t ube, and f lickr. *Journal of the Association for Information Science and Technology*, 66(10):2045–2060, 2015.

[Ohlsson 2011] Stellan Ohlsson. *Deep learning: How the mind overrides experience*. Cambridge University Press, 2011.

[Okuno *et al.* 2014a] Syunya Okuno, Hiroki Asai, and Hayato Yamana. A challenge of authorship identification for ten-thousand-scale microblog users. In *2014 IEEE international conference on big data (big data)*, pages 52–54. IEEE, 2014.

[Okuno *et al.* 2014b] Syunya Okuno, Hiroki Asai, and Hayato Yamana. A challenge of authorship identification for ten-thousand-scale microblog users. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 52–54, 2014.

[Olaniyan and Akpojivi 2021] Akintola Olaniyan and Ufuoma Akpojivi. Transforming communication, social media, counter-hegemony and the struggle for the soul of nigeria. *Information, Communication & Society*, 24(3):422–437, 2021.

[Omar 2018] Abdul-Majeed Omar. Psychology of language acquisition and EFL teaching methodology: A critical overview. *Gezira Journal of Educational Sciences and Humanities*, 12(2), 2018.

[O'neil 2014] Patrick O'neil. *DATABASE: principles programming performance*. Morgan Kaufmann, 2014.

[Ong *et al.* 2009] Chorng-Shyong Ong, Min-Yuh Day, and Wen-Lian Hsu. The measurement of user satisfaction with question answering systems. *Information & Management*, 46(7):397–403, 2009.

[OpenNLP 2011] Apache OpenNLP. Apache software foundation. *URL http://opennlp.apache.org*, 2011.

[Oram 2001] Peter Oram. Wordnet: An electronic lexical database. christiane fellbaum (ed.). cambridge, ma: Mit press, 1998. pp. 423. *Applied Psycholinguistics*, 22(1):131–134, 2001.

[Orikogbo *et al.* 2016] Damilola Orikogbo, Matthias Büchler, and Manuel Egele. Crios: toward large-scale iOS application analysis. In *Proceedings of the 6th Workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 33–42. ACM, 2016.

[Ortega-Mendoza *et al.* 2018] Rosa María Ortega-Mendoza, A Pastor López-Monroy, Anilu Franco-Arcega, and Manuel Montes-y Gómez. Emphasizing personal information for author profiling: New approaches for term selection and weighting. *Knowledge-Based Systems*, 145:169–181, 2018.

[Ouamour and Sayoud 2013] Siham Ouamour and Halim Sayoud. Authorship attribution of short historical arabic texts based on lexical features. In *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 144–147. IEEE, 2013.

[Oussalah *et al.* 2013] M. Oussalah, F. Bhat, K. Challis, and T. Schnier. A software architecture for twitter collection, search and geolocation services. *Knowledge-Based Systems*, 37:105–120, 2013.

[Padda 2020] Karmvir Padda. Fake news on twitter in 2016 us presidential election: a quantitative approach. *The Journal of Intelligence, Conflict, and Warfare*, 3(2):1–24, 2020.

[Paltridge 2012] Brian Paltridge. Genre and english for specific purposes. *The handbook of English for specific purposes*, pages 347–366, 2012.

[Pang and Lee 2005] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACLâ€™05)*, pages 115–124, 2005.

[Papangelis *et al.* 2017] Alexandros Papangelis, Panagiotis Papadakos, Margarita Kotti, Yannis Stylianou, Yannis Tzitzikas, and Dimitris Plexousakis. Ld-sds: Towards an expressive spoken dialogue system based on linked-data. *arXiv preprint arXiv:1710.02973*, 2017.

[Papantoniou and Tzitzikas 2019] Katerina Papantoniou and Yannis Tzitzikas. Cs563-qa: A collection for evaluating question answering systems. *arXiv preprint arXiv:1907.01611*, 2019.

[Parikh *et al.*] Kieran Sagar Parikh, Vinodini Venkataram, and Jugal Kalita. Towards a universal document encoder for authorship attribution.

[Park *et al.* 2016] Andrew J Park, Brian Beck, Darrick Fletche, Patrick Lam, and Herbert H Tsang. Temporal analysis of radical dark web forum users. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 880–883. IEEE, 2016.

[Pasieka *et al.* 2022] Nadiia Pasieka, Vasyl Sheketa, Myroslava Kulynych, Yulia Romanyshyn, Svitlana Chupakhina, and Olha Khytrova. Machine training for intelligent analysis of text for the identification of the author. 2022.

[Pasquini *et al.* 2021] Cecilia Pasquini, Irene Amerini, and Giulia Boato. Media forensics on social media platforms: a survey. *EURASIP Journal on Information Security*, 2021(1):4, 2021.

[Pass 2007] Susan Pass. When constructivists Jean Piaget and Lev Vygotsky were pedagogical collaborators: A viewpoint from a study of their communications. *Journal of Constructivist Psychology*, 20(3):277–282, 2007.

[Pazos R *et al.* 2013] Rodolfo A Pazos R, Juan J González B, Marco A Aguirre L, José A Martínez F, and Héctor J Fraire H. Natural language interfaces to databases: an analysis of the state of the art. *Recent Advances on Hybrid Intelligent Systems*, pages 463–480, 2013.

[Pazos Rangel *et al.* 2014] Rodolfo A Pazos Rangel, Marco A Aguirre, Juan J González, and Juan Martín Carpio. Features and pitfalls that users should seek in natural language interfaces to databases. *Recent Advances on Hybrid Approaches for Designing Intelligent Systems*, pages 617–630, 2014.

[Peersman *et al.* 2011] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44, 2011.

[Peirce 1902] Charles Sanders Peirce. Logic as semiotic: The theory of signs. *Philosophical writings of Peirce*, page 100, 1902.

[Peñas and Rodrigo 2011] Anselmo Peñas and Alvaro Rodrigo. A simple measure to assess non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1415–1424. Association for Computational Linguistics, 2011.

[Peng *et al.* 2003a] Fuchun Peng, Dale Schuurmans, Vlado Keselj, and Shaojun Wang. Language independent authorship attribution with character level n-grams. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, 2003.

[Peng *et al.* 2003b] Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. Language independent authorship attribution using character level language models. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, page 267â€"274, USA, 2003. Association for Computational Linguistics.

[Peng *et al.* 2004] Fuchun Peng, Dale Schuurmans, and Shaojun Wang. Augmenting Naive Bayes Classifiers with Statistical Language Models. *Information Retrieval*, 7(3):317–345, 2004.

[Peng *et al.* 2015] Nanyun Peng, Francis Ferraro, Mo Yu, Nicholas Andrews, Jay DeYoung, Max Thomas, Matthew R Gormley, Travis Wolfe, Craig Harman, Benjamin Van Durme, et al. A concrete Chinese NLP pipeline. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 86–90, 2015.

[Peng *et al.* 2018a] Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, 2018.

[Peng *et al.* 2018b] Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Pennacchiotti and Popescu 2011] Marco Pennacchiotti and Ana-Maria Popescu. A machine learning approach to twitter user classification. In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 281–288, 2011.

[Pennebaker *et al.* 2001] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[Pennebaker 2011] James W Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

[Pennington *et al.* 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Pennington 1987] Nancy Pennington. Stimulus structures and mental representations in expert comprehension of computer programs. *Cognitive psychology*, 19(3):295–341, 1987.

[Perer and Shneiderman 2008] Adam Perer and Ben Shneiderman. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 265â€"274, New York, NY, USA, 2008. Association for Computing Machinery.

[Pérez-Sabater 2012] Carmen Pérez-Sabater. The linguistics of social networking: A study of writing conventions on facebook. *Linguistik online*, 56(6), 2012.

[Perito *et al.* 2011] Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. How unique and traceable are usernames? In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 1–17. Springer, 2011.

[Perrin 2003] Dominique Perrin. Automata and formal languages. *Université de Marne-la-vallée*, pages 1–22, 2003.

[Perry 2016] Brain Perry. Advanced guide to the Bloomberg terminal: Bond market functions. *Investopedia Online*, 2016.

[Peterson *et al.* 2011] Kelly Peterson, Matt Hohensee, and Fei Xia. Email formality in the workplace: A case study on the enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, 2011.

[Petre and Blackwell 2007] Marian Petre and Alan F Blackwell. Children as unwitting end-user programmers. In *IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 239–242. IEEE, 2007.

[Pfeffer *et al.* 2023] Juergen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, et al. Just another day on twitter: a complete 24 hours of twitter data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1073–1081, 2023.

[Pickard and others 2007] Mary J Pickard et al. The new Bloomâ€™s taxonomy: An overview for family and consumer sciences. *Journal of Family and Consumer Sciences Education*, 25(1), 2007.

[Pillay and Solorio 2010] Sangita R. Pillay and Thamar Solorio. Authorship attribution of web forum posts. In *2010 eCrime Researchers Summit*, pages 1–7, Oct 2010.

[Pillet and Carillo 2016] Jean-Charles Pillet and Kevin Daniel AndrÃ© Carillo. Email-free collaboration: An exploratory study on the formation of new work habits among knowledge workers. *International Journal of Information Management*, 36(1):113–125, 2016.

[Pinto *et al.* 2014] David Pinto, Helena GÃ³mez-Adorno, Darnes VilariÃ±o, and Vivek Kumar Singh. A graph-based multi-level linguistic representation for document understanding. *Pattern Recognition Letters*, 41:93–102, 2014. Supervised and Unsupervised Classification Techniques and their Applications.

[Pisanski 2014] Katarzyna Pisanski. *Human vocal communication of body size*. PhD thesis, 2014.

[Piwek and Joinson 2016] Lukasz Piwek and Adam Joinson. "what do they snapchat about?" patterns of use in time-limited instant messaging service. *Computers in Human Behavior*, 54:358–367, 2016.

[Plakias and Stamatatos 2008] Spyridon Plakias and Efstathios Stamatatos. Tensor space models for authorship identification. In *Hellenic Conference on Artificial Intelligence*, pages 239–249. Springer, 2008.

[Pokou *et al.* 2016] Yao Jean Marc Pokou, Philippe Fournier-Viger, and Chadia Moghrabi. Authorship attribution using small sets of frequent part-of-speech skip-grams. In *The Twenty-Ninth International Flairs Conference*, 2016.

[Politis 2008] Diamanto Politis. The process of entrepreneurial learning: a conceptual framework. In *Entrepreneurial Learning*, pages 66–93. Routledge, 2008.

[Pons *et al.* 2016] Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343, 2016.

[Ponto 2015] Julie Ponto. Understanding and evaluating survey research. *Journal of the advanced practitioner in oncology*, 6(2):168, 2015.

[Poria *et al.* 2016] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59, 2016.

[Posadas-Durán *et al.* 2015] Juan-Pablo Posadas-Durán, Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, Alexander Gelbukh, and Obdulia Pichardo-Lagunas. Syntactic n-grams as features for the author profiling task. *Working Notes Papers of the CLEF*, 2015.

[Posadas-Durán *et al.* 2017] Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, David Pinto, and Liliana Chanona-Hernández. Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing*, 21(3):627–639, 2017.

[Post 1944] Emil L Post. Recursively enumerable sets of positive integers and their decision problems. *bulletin of the American Mathematical Society*, 50(5):284–316, 1944.

[Potha and Stamatatos 2014] Nektaria Potha and Efstathios Stamatatos. A profile-based method for authorship verification. In *Hellenic Conference on Artificial Intelligence*, pages 313–326. Springer, 2014.

[Potthast *et al.* 2016] Martin Potthast, Matthias Hagen, and Benno Stein. Author obfuscation: Attacking the state of the art in authorship verification. *CLEF (Working Notes)*, pages 716–749, 2016.

[Powell and Wimmer 2015] Loreen M Powell and Hayden Wimmer. Evaluating the effectiveness of self-created student screencasts as a tool to increase student learning outcomes in a hands-on computer programming course. *Information Systems Education Journal*, 13(5):106, 2015.

[Prabhumoye *et al.* 2018] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*, 2018.

[Prasad *et al.* 2015] S. Naga Prasad, V.B. Narsimha, P. Vijayapal Reddy, and A. Vinaya Babu. Influence of lexical, syntactic and structural features and their combination on authorship attribution for telugu text. *Procedia Computer Science*, 48:58–64, 2015. International Conference on Computer, Communication and Convergence (ICCC 2015).

[Preotiuc-Pietro *et al.* 2019] Daniel Preotiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. Automatically identifying complaints in social media. *CoRR*, abs/1906.03890, 2019.

[Press and Wilson 1978] S James Press and Sandra Wilson. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364):699–705, 1978.

[Price 2004] Linda Price. Individual differences in learning: Cognitive control, cognitive style, and learning style. *Educational Psychology*, 24(5):681–698, 2004.

[Prior and Lister 2004] Julia Coleman Prior and Raymond Lister. The backwash effect on SQL skills grading. *ACM Intelligent Systems and Informatics Bulletin*, 36(3):32–36, 2004.

[Prior 2003] Julia Coleman Prior. Online assessment of SQL query formulation skills. In *Proceedings of the fifth Australasian conference on Computing education-Volume 20*, pages 247–256. Australian Computer Society, Inc., 2003.

[Prior 2014] Julia R Prior. AsseSQL: an online, browser-based SQL skills assessment tool. In *Proceedings of the Innovation & technology in computer science education conference*, pages 327–327. ACM, 2014.

[Pritchard 2017] Alan Pritchard. *Ways of learning: Learning theories for the classroom*. Routledge, 2017.

[Proisl and Uhrig 2016] Thomas Proisl and Peter Uhrig. Somajo: State-of-the-art tokenization for german web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop*, pages 57–62, 2016.

[Pudaruth *et al.* 2014] Sameerchand Pudaruth, Sandiana Amourdon, and Joey Anseline. Automated generation of song lyrics using CFGs. In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pages 613–616. IEEE, 2014.

[Puschmann 2013] Cornelius Puschmann. 4. blogging. *Pragmatics of computer-mediated communication*, pages 83–108, 2013.

[Pyott and Sanders 1991] Sean Pyott and Ian Sanders. Alex: an aid to teaching algorithms. *ACM SIGCSE Bulletin*, 23(3):36–44, 1991.

[Qi and Davison 2009] Xiaoguang Qi and Brian D Davison. Web page classification: Features and algorithms. *ACM computing surveys (CSUR)*, 41(2):12, 2009.

[Qian *et al.* 2014] Tieyun Qian, Bing Liu, Li Chen, and Zhiyong Peng. Tri-training for authorship attribution with limited training data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 345–351, 2014.

[Qian *et al.* 2015] Tie-Yun Qian, Bing Liu, Qing Li, and Jianfeng Si. Review authorship attribution in a similarity space. *Journal of Computer Science and Technology*, 30(1):200–213, 2015.

[Qian 2012] Gang Qian. Designing and implementing unsupervised online database labs. *Journal of Computing Sciences in Colleges*, 27(4):30–36, 2012.

[Qin *et al.* 2018] Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. DeepEye: Visualizing your data by keyword search. In *EDBT*, pages 441–444, 2018.

[Quandt *et al.* 2022] Thorsten Quandt, Johanna Klapproth, and Lena Frischlich. Dark social media participation and well-being. *Current Opinion in Psychology*, 45:101284, 2022.

[Quinlan 1986] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[Quinlan 2014] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[Quiring *et al.* 2019] Erwin Quiring, Alwin Maier, and Konrad Rieck. Misleading authorship attribution of source code using adversarial learning. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 479–496, 2019.

[Ragel *et al.* 2013] Roshan Ragel, Pramod Herath, and Upul Senanayake. Authorship detection of sms messages using unigrams. In *2013 IEEE 8th International Conference on Industrial and Information Systems*, pages 387–392. IEEE, 2013.

[Raghavan *et al.* 2010] Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 conference short papers*, pages 38–42, 2010.

[Raja 2014] Rao DV Raja. Method and system for automatic switching between chat windows. *Google Patents, US Patent App. 13/738,785*, 2014.

[Rajpurkar *et al.* 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[Ramage *et al.* 2009] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 248–256, 2009.

[Ramalingam and Wiedenbeck 1997] Vennila Ramalingam and Susan Wiedenbeck. An empirical study of novice program comprehension in the imperative and object-oriented styles. In *the Seventh Workshop on Empirical studies of programmers*, pages 124–139. ACM, 1997.

[Ramírez Sánchez *et al.* 2021] Julián Ramírez Sánchez, Alejandra Campo-Archbold, Andrés Zapata Rozo, Daniel Díaz-López, Javier Pastor-Galindo, Félix Gómez Mármol, and Julián Aponte Díaz. Uncovering cybercrimes in social media through natural language processing. *Complexity*, 2021:1–15, 2021.

[Ramnial *et al.* 2016a] Hoshiladevi Ramnial, Shireen Panchoo, and Sameerchand Pudaruth. Authorship attribution using stylometry and machine learning techniques. In *Intelligent Systems Technologies and Applications*, pages 113–125. Springer, 2016.

[Ramnial *et al.* 2016b] Hoshiladevi Ramnial, Shireen Panchoo, and Sameerchand Pudaruth. Gender profiling from phd theses using k-nearest neighbour and sequential minimal optimisation. In *Intelligent Systems Technologies and Applications*, pages 369–377. Springer, 2016.

[Rashid *et al.* 2013a] Awais Rashid, Alistair Baron, Paul Rayson, Corinne May-Chahal, Phil Greenwood, and James Walkerdine. Who am i? analyzing digital personas in cybercrime investigations. *Computer*, 46(4):54–61, 2013.

[Rashid *et al.* 2013b] Awais Rashid, Alistair Baron, Paul Rayson, Corinne May-Chahal, Phil Greenwood, and James Walkerdine. Who am i? analyzing digital personas in cybercrime investigations. *Computer*, 46(4):54–61, April 2013.

[Rathinam and Sivakumar 2011] Ashok Prasanna Rathinam and Ramani Sivakumar. *System and Method for Managing One or More Databases*, 2011. US Patent App. 12/634,844.

[Rayana and Akoglu 2015] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994. ACM, 2015.

[Reavley and Pilkington 2014] Nicola J. Reavley and Pamela D. Pilkington. Use of twitter to monitor attitudes toward depression and schizophrenia: An exploratory study. *PeerJ*, 2014:e647, 10 2014.

[Reddy *et al.* 2016] T Raghunadha Reddy, B Vishnu Vardhan, and P Vijaypal Reddy. A survey on authorship profiling techniques. *International Journal of Applied Engineering Research*, 11(5):3092–3102, 2016.

[Reinaldha and Widagdo 2014] Filbert Reinaldha and Tricya E Widagdo. Natural language interfaces to database (NLIDB): Question handling and unit conversion. In *2014 International Conference on Data and Software Engineering (ICODSE)*, pages 1–6. IEEE, 2014.

[Reips and Garaizar 2011] Ulf-Dietrich Reips and Pablo Garaizar. Mining twitter: A source for psychological wisdom of the crowds. *Behavior research methods*, 43(3):635–642, 2011.

[Reisner 1977] Phyllis Reisner. Use of psychological experimentation as an aid to development of a query language. *IEEE Transactions on Software Engineering*, (3):218–229, 1977.

[Reiter and Dale 2000] Ehud Reiter and Robert Dale. *Building natural language generation systems*. Cambridge university press, 2000.

[Renaud and Van Biljon 2004] Karen Renaud and Judy Van Biljon. Teaching SQL - Which Pedagogical Horse for This Course? In *British National Conference on Databases*, pages 244–256. Springer, 2004.

[Renaud *et al.* 2009] K Renaud, H Al Shuailu, and R Cooper. Facilitating efficacious transfer of database knowledge and skills. In *Ideas in Teaching, Learning and Assessment of Databases: A Communication of the 7 th International Workshop on Teaching, Learning and Assessment of Databases (TLAD 2009)*, pages 213–224. Springer, 2009.

[Renkl 2014] Alexander Renkl. Toward an instructionally oriented theory of example-based learning. *Cognitive science*, 38(1):1–37, 2014.

[Resende de Mendonça *et al.* 2020] Ricardo Resende de Mendonça, Daniel Felix de Brito, Ferrucio de Franco Rosa, Júlio Cesar dos Reis, and Rodrigo Bonacin. A framework for detecting intentions of criminal acts in social media: A case study on twitter. *Information*, 11(3):154, 2020.

[Resnick *et al.* 2009] Mitchel Resnick, John Maloney, Andrés Monroy-Hernández, Natalie Rusk, Evelyn Eastmond, Karen Brennan, Amon Millner, Eric Rosenbaum, Jay Silver, Brian Silverman, et al. Scratch: programming for all. *Communications of the ACM*, 52(11):60–67, 2009.

[Resnick 2013] Mitchel Resnick. Learn to code, code to learn. *EdSurge, May*, 54, 2013.

[Rezaeinia *et al.* 2019] Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:139–147, 2019.

[Ribeiro *et al.* 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135â€“1144, New York, NY, USA, 2016. Association for Computing Machinery.

[Ringnér 2008] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

[Rish and others 2001] Irina Rish et al. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[Ritchie *et al.* 1992] Graeme D Ritchie, Graham J Russell, Alan W Black, and Stephen G Pulman. *Computational morphology: practical mechanisms for the English lexicon*. MIT press, 1992.

[Rizvi *et al.* 2011] Mona Rizvi, Thorna Humphries, Debra Major, Meghan Jones, and Heather Lauzun. A CS0 course using Scratch. *Journal of Computing Sciences in Colleges*, 26(3):19–27, 2011.

[Roark and Johnson 1999] Brian Roark and Mark Johnson. Efficient probabilistic top-down and left-corner parsing. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 421–428. Association for Computational Linguistics, 1999.

[Roberts *et al.* 2019] Andrew Fiske Roberts, Craig W Stanfill, and Adam Harris Weiss. *Computer screen with visual programming icons*, 2019. US Patent App. 29/582,279.

[Robins *et al.* 2003] Anthony Robins, Janet Rountree, and Nathan Rountree. Learning and teaching programming: A review and discussion. *Computer Science Education*, 13(2):137–172, 2003.

[Robnik-Šikonja 2004] Marko Robnik-Šikonja. Improving random forests. In *Machine Learning: ECML 2004: 15th European Conference on Machine Learning, Pisa, Italy, September 20-24, 2004. Proceedings 15*, pages 359–370. Springer, 2004.

[Rocha *et al.* 2016] Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. Authorship attribution for social media forensics. *IEEE transactions on information forensics and security*, 12(1):5–33, 2016.

[Rocha *et al.* 2021a] Yasmim Mendes Rocha, Gabriel AcÃ!'cio de Moura, Gabriel Alves DesidÃ©rio, Carlos Henrique de Oliveira, Francisco Dantas LourenÃ§o, and Larissa Deadame de Figueiredo Nicolete. The impact of fake news on social media and its influence on health during the covid-19 pandemic: a systematic review. *Journal of Public Health (Germany)*, pages 1–10, 10 2021.

[Rocha *et al.* 2021b] Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review. *Journal of Public Health*, pages 1–10, 2021.

[Roche and Schabes 1997] Emmanuel Roche and Yves Schabes. *Finite-state language processing*. MIT press, 1997.

[Rockoff 2016] Larry Rockoff. *The language of SQL*. Addison-Wesley Professional, 2016.

[Rodrigues *et al.* 2018] Ricardo Rodrigues, Hugo Gonçalo Oliveira, and Paulo Gomes. NLP-port: A pipeline for portuguese NLP (short paper). In *7th Symposium on Languages, Applications and Technologies (SLATE 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[Rodríguez *et al.* 2016] Rosa M Rodríguez, B Bedregal, Humberto Bustince, YC Dong, Bahram Farhadinia, Cengiz Kahraman, L Martínez, Vicenç Torra, YJ Xu, ZS Xu, et al. A position and perspective analysis of hesitant fuzzy sets on information fusion in decision making. towards high quality progress. *Information Fusion*, 29:89–97, 2016.

[Rogati and Yang 2002] Monica Rogati and Yiming Yang. High-performing feature selection for text classification. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, page 659â€"661, New York, NY, USA, 2002. Association for Computing Machinery.

[Rojit *et al.* 2016] Vidhu Rojit, Sindhu R Pai, Shruthi Kaivalya, and Viraj Kumar. Visual specifications for web-application programming assignments. In *2016 IEEE Eighth International Conference on Technology for Education (T4E)*, pages 50–53. IEEE, 2016.

[Rong 2014] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[Room 2019] Chat Room. Natural language processing. 2019.

[Rosenberg *et al.* 2020] Hans Rosenberg, Shahbaz Syed, and Salim Rezaie. The twitter pandemic: The critical role of twitter in the dissemination of medical information and misinformation during the covid-19 pandemic. *Canadian journal of emergency medicine*, 22(4):418–421, 2020.

[Rosenfeld *et al.* 2017] Ariel Rosenfeld, Abejide Ade-Ibijola, and Sigrid Ewert. Regex parser ii: teaching regular expression fundamentals via educational gaming. In *Annual Conference of the Southern African Computer Lecturers' Association*, pages 99–112. Springer, 2017.

[Rosenfeld 2000] Ronald Rosenfeld. Two decades of statistical language modeling: Where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278, 2000.

[Roy *et al.* 2019] Pradeep Kumar Roy, Jyoti Prakash Singh, and Snehasish Banerjee. Deep learning to filter SMS spam. *Future Generation Computer Systems*, 2019.

[Ruan *et al.* 2018] Wei Ruan, Naveenkumar Appasani, Katherine Kim, Joseph Vincelli, Hyun Kim, and Won-Sook Lee. Pictorial visualization of emr summary interface and medical information extraction of clinical notes. In *2018 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pages 1–6. IEEE, 2018.

[Ruder *et al.* 2016a] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint arXiv:1609.06686*, 2016.

[Ruder *et al.* 2016b] Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. *Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution*, 2016.

[Ruder 2019] Sebastian Ruder. *Neural transfer learning for natural language processing*. PhD thesis, NUI Galway, 2019.

[Rudman 1997] Joseph Rudman. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365, 1997.

[Rudman 2012] Joseph Rudman. The state of non-traditional authorship attribution studiesâ€"2012: Some problems and solutions. *English Studies*, 93(3):259–274, 2012.

[Ruohonen 2009] Keijo Ruohonen. Formal languages. *Lecture Notes*, pages 1–94, 2009.

[Rush 2018] Alexander M Rush. The annotated transformer. In *Proceedings of workshop for NLP open source software (NLP-OSS)*, pages 52–60, 2018.

[Russell-Rose *et al.* 2002] Tony Russell-Rose, Mark Stevenson, and Miles Whitehead. The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. 2002.

[Rutten *et al.* 2012] Nico Rutten, Wouter R Van Joolingen, and Jan T Van der Veen. The learning effects of computer simulations in science education. *Computers & Education*, 58(1):136–153, 2012.

[Ryan *et al.* 2015] James Owen Ryan, Eric Kaltman, Michael Mateas, and Noah Wardrip-Fruin. What we talk about when we talk about games: Bottom-up game studies using natural language processing. *Proc. FDG*, 2015.

[Ryan *et al.* 2017] Tracii Ryan, Kelly A Allen, DeLeon L Gray, and Dennis M McInerney. How social are social media? a review of online social behaviour and connectedness. *Journal of Relationships Research*, 8:e8, 2017.

[Saavedra *et al.* 2011] Serguei Saavedra, Kathleen Hagerty, and Brian Uzzi. Synchronicity, instant messaging, and performance among financial traders. *Proceedings of the National Academy of Sciences*, 108(13):5296–5301, 2011.

[Sacerdoti 1977] Earl D Sacerdoti. *Language access to distributed data with error recovery*. Technical report, SRI International Menlo Park CA, Artificial Intelligence Center, 1977.

[Sadiq *et al.* 2004] Shazia Sadiq, Maria Orlowska, Wasim Sadiq, and Joe Lin. SQLator: an online SQL learning workbench. In *ACM Special Interest Group on Computer Science Education Bulletin*, volume 36, pages 223–227. ACM, 2004.

[Safavian and Landgrebe 1991] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[Sag *et al.* 2002] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer, 2002.

[Sagi and Rokach 2018] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[Sagiroglu and Sinanc 2013] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In *International Conference on Collaboration Technologies and Systems*, pages 42–47. IEEE, 2013.

[Saha *et al.* 2016] Diptikalyan Saha, Avrilia Floratou, Karthik Sankaranarayanan, Umar Farooq Minhas, Ashish R Mittal, and Fatma Özcan. Athena: an ontology-driven system for natural language querying over relational data stores. *Proceedings of the VLDB Endowment*, 9(12):1209–1220, 2016.

[Saha *et al.* 2019] Sajal Saha, Mazid Alam, and Smita Dey. A framework for artificially intelligent customized voice response system design using speech synthesis markup. *Intelligent Speech Signal Processing*, page 175, 2019.

[Sailunaz *et al.* 2018] Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26, 2018.

[Saini and Rakholia 2016] Jatinderkumar R Saini and Rajnish M Rakholia. On continent and script-wise divisions-based statistical measures for stop-words lists of international languages. *Procedia Computer Science*, 89:313–319, 2016.

[Saini *et al.* 2019] Mayank Saini, Sharad Verma, and Aditi Sharan. Multi-view ensemble learning using rough set based feature ranking for opinion spam detection. In *Advances in Computer Communication and Computational Sciences*, pages 3–12. Springer, 2019.

[Sakaki *et al.* 2011] Takeshi Sakaki, Fujio Toriumi, and Yutaka Matsuo. Tweet trend analysis in an emergency situation. In *Proceedings of the Special Workshop on Internet and Disasters*, page 3. ACM, 2011.

[Sakaki *et al.* 2013] Takeshi Sakaki, Masahide Okazaki, and Yoshikazu Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931, 2013.

[Salloum *et al.* 2017] Said A Salloum, Mostafa Al-Emran, Azza Abdel Monem, and Khaled Shaalan. A survey of text mining in social media: Facebook and twitter perspectives. *Advances in Science, Technology and Engineering Systems Journal*, 2(1):127–133, 2017.

[Salmivalli 2010] Christina Salmivalli. Bullying and the peer group: A review. *Aggression and Violent Behavior*, 15(2):112–120, 2010. Special Issue on Group Processes and Aggression.

[Sampasa-Kanyinga *et al.* 2014] Hugues Sampasa-Kanyinga, Paul Roumeliotis, and Hao Xu. Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among canadian schoolchildren. *PloS one*, 9(7):e102145, 2014.

[Sánchez-Moya and Cruz-Moya 2015] Alfonso Sánchez-Moya and Olga Cruz-Moya. Whatsapp, textese, and moral panics: discourse features and habits across two generations. *Procedia-Social and Behavioral Sciences*, 173:300–306, 2015.

[Sander and Wauer 2019] André Sander and Roland Wauer. Integrating terminologies into standard SQL: a new approach for research on routine data. *Journal of biomedical semantics*, 10(1):7, 2019.

[Sanderson and Guenter 2006] Conrad Sanderson and Simon Guenter. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 482–491. Association for Computational Linguistics, 2006.

[Sandoz *et al.* 2018] Romain Sandoz, Sarah Composto, Sandrine Divorne, Olivier Ertz, and Jens Ingensand. *GeoSQL Journey-A gamified learning experience to introduce (or demystify) geospatial SQL queries*. Technical report, PeerJ Preprints, 2018.

[Santini 2006] Marina Santini. Web pages, text types, and linguistic features: Some issues. *ICAME journal*, 30:67–86, 2006.

[Sapkota *et al.* 2014] Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. Cross-topic authorship attribution: Will out-of-topic data help? In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1228–1237, 2014.

[Sapkota *et al.* 2015] Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 93–102, 2015.

[Sarfraz 2017] Haniya Sarfraz. Strategic leadership development: simplified with Bloomâ€™s taxonomy. *Industrial and Commercial Training*, 49(1):40–47, 2017.

[Sari *et al.* 2017] Yunita Sari, Andreas Vlachos, and Mark Stevenson. Continuous n-gram representations for authorship attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 267–273, Valencia, Spain, April 2017. Association for Computational Linguistics.

[Sari *et al.* 2018a] Yunita Sari, Mark Stevenson, and Andreas Vlachos. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[Sari *et al.* 2018b] Yunita Sari, Mark Stevenson, and Andreas Vlachos. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[Sari 2018] Yunita Sari. *Neural and Non-neural Approaches to Authorship Attribution*. PhD thesis, University of Sheffield, 2018.

[Sarwar *et al.* 2018a] Raheem Sarwar, Chenyun Yu, Sarana Nutanong, Norawit Urailertprasert, Nattapol Vannaboot, and Thanawin Rakthanmanon. A scalable framework for stylometric analysis of multi-author documents. In *International Conference on Database Systems for Advanced Applications*, pages 813–829. Springer, 2018.

[Sarwar *et al.* 2018b] Raheem Sarwar, Chenyun Yu, Ninad Tungare, Kanatip Chitavisutthivong, Sukrit Sriratanawilai, Yaohai Xu, Dickson Chow, Thanawin Rakthanmanon, and Sarana Nutanong. An effective and scalable framework for authorship attribution query processing. *IEEE Access*, 6:50030–50048, 2018.

[Satyanarayan and Heer 2014] Arvind Satyanarayan and Jeffrey Heer. Lyra: An interactive visualization design environment. In *Computer Graphics Forum*, volume 33, pages 351–360. Wiley Online Library, 2014.

[Savage 2017] Ron Savage. *Database Language SQL (SQL-2003) SQL/Foundation*. https://ronsavage.github.io/SQL/sql-2003-2.bnf.html, 2017. Accessed: 2018-09-30.

[Savoy 2012] Jacques Savoy. Authorship attribution based on specific vocabulary. *ACM Trans. Inf. Syst.*, 30(2), may 2012.

[Savoy 2013] Jacques Savoy. Authorship attribution based on a probabilistic topic model. *Information Processing & Management*, 49(1):341–354, 2013.

[Schaetti 2015] Nils Schaetti. Character-based convolutional neural network for style change detection. *Training*, 2980(1490):1490, 2015.

[Schaetti 2017] Nils Schaetti. Unine at clef 2017: Tf-idf and deep-learning for author profiling. In *CLEF (Working Notes)*, 2017.

[Schlager and Ogden 1986] MS Schlager and William C Ogden. A cognitive model of database querying: a tool for novice instruction. *ACM SIGCHI Bulletin*, 17(4):107–113, 1986.

[Schler *et al.* 2006a] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[Schler *et al.* 2006b] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205, 2006.

[Schmitz 2007] Sylvain Schmitz. *Approximating context-free grammars for parsing and verification*. PhD thesis, Université Nice Sophia Antipolis, 2007.

[Schölkopf *et al.* 2002] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[Schulte *et al.* 2010] Carsten Schulte, Tony Clear, Ahmad Taherkhani, Teresa Busjahn, and James H Paterson. An introduction to program comprehension for computer science educators. In *Proceedings of the 2010 ITiCSE working group reports*, pages 65–86. ACM, 2010.

[Schuster and Paliwal 1997] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[Schwartz *et al.* 2013a] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[Schwartz *et al.* 2013b] Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on empirical methods in natural language processing*, pages 1880–1891, 2013.

[Schweinsberg and Wegner 2017] Kai Schweinsberg and Lutz Wegner. Advantages of complex SQL types in storing XML documents. *Future Generation Computer Systems*, 68:500–507, 2017.

[Scott III 2010] Robert H Scott III. Bloomberg 101. *Journal of Financial Education*, pages 80–88, 2010.

[Sebastiani 2002] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[Sebastiani 2005] Fabrizio Sebastiani. Text categorization. In *Encyclopedia of Database Technologies and Applications*, pages 683–687. IGI Global, 2005.

[Sebastiani 2006] Fabrizio Sebastiani. Classification of text, automatic. *The encyclopedia of language and linguistics*, 14:457–462, 2006.

[Seddighi *et al.* 2020] Hamed Seddighi, Ibrahim Salmani, and Saeideh Seddighi. Saving lives and changing minds with twitter in disasters and pandemics: a literature review. *Journalism and Media*, 1(1):59–77, 2020.

[Seffah 2015] Ahmed Seffah. *Patterns of HCI design and HCI design of patterns: bridging HCI design and model-driven software engineering*. Springer, 2015.

[Segarra *et al.* 2015] Santiago Segarra, Mark Eisen, and Alejandro Ribeiro. Authorship attribution through function word adjacency networks. *IEEE Transactions on Signal Processing*, 63(20):5464–5478, 2015.

[Seidman 2013] Shachar Seidman. Authorship verification using the impostors method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*. Citeseer, 2013.

[Seidmann 2021] Vered Seidmann. On blogs, autistic bloggers, and autistic space. *Information, Communication & Society*, 24(15):2277–2292, 2021.

[Seligman 1970] Martin E Seligman. On the generality of the laws of learning. *Psychological review*, 77(5):406, 1970.

[Selvi *et al.* 2017] S Thamarai Selvi, P Karthikeyan, A Vincent, V Abinaya, G Neeraja, and R Deepika. Text categorization using rocchio algorithm and random forest algorithm. In *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pages 7–12. IEEE, 2017.

[Semino and Culpeper 2002] Elena Semino and Jonathan Culpeper. *Cognitive stylistics: Language and cognition in text analysis*, volume 1. John Benjamins Publishing, 2002.

[Sengupta and Garg 2019] Subhasree Sengupta and Radhika Garg. Impact of voice-based interaction on learning practices and behavior of children. In *Joint Proceedings of the ACM IUI Workshops, Los Angeles, USA, March 20*, pages 1–3, 2019.

[Sennrich *et al.* 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.

[Serban *et al.* 2016a] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Serban *et al.* 2016b] Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*, 2016.

[Seroussi *et al.* 2010] Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. Collaborative inference of sentiments from texts. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 195–206. Springer, 2010.

[Seroussi *et al.* 2011] Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. Authorship attribution with latent dirichlet allocation. In *Proceedings of the fifteenth conference on computational natural language learning*, pages 181–189. Association for Computational Linguistics, 2011.

[Seroussi *et al.* 2014] Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310, 2014.

[Seuren 2017] Pieter Seuren. *Semantic syntax*. Brill, 2017.

[Seyed-Abbassi 1993] Behrooz Seyed-Abbassi. A SQL project as a learning method in a database course. In *Proceedings of the 1993 conference on Computer personnel research*, pages 291–297. ACM, 1993.

[Seyyedi and Minaei-Bidgoli 2018] Seyyed Hossein Seyyedi and Behrouz Minaei-Bidgoli. Estimator learning automata for feature subset selection in high-dimensional spaces, case study: Email spam detection. *International Journal of Communication Systems*, 31(8):e3541, 2018.

[Shabaz *et al.* 2015] K Shabaz, Jim D O'Shea, Keeley A Crockett, and A Latham. Aneesah: A conversational natural language interface to databases. In *Proceedings of the World Congress on Engineering*, volume 1, 2015.

[Shackelford 1997] Russell L Shackelford. *Introduction to computing and algorithms*. Addison-Wesley Longman Publishing Co., Inc., 1997.

[Shaffer 2020] Ryan Shaffer. *Technology's Impact on War and Terrorism: Terrorism in Cyberspace: The Next Generation, by Gabriel Weimann, Washington DC/New York, Woodrow Wilson Center Press/Columbia University Press, 2015, xvii+ 296 pp., 30.00 (paperback)*, 2020.

[Shah and Sengupta 2018] Hirav Shah and Amit Sengupta. Designing mobile based computational support for low-literate community health workers. *International Journal of Human-Computer Studies*, 115:1–8, 2018.

[Shah *et al.* 2013] Axita Shah, Jyoti Pareek, Hemal Patel, and Namrata Panchal. NLKBIDB-natural language and keyword based interface to database. In *2013 International conference on advances in computing, communications and informatics (ICACCI)*, pages 1569–1576. IEEE, 2013.

[Shah *et al.* 2017] Philipp Shah, Marc Berges, and Peter Hubwieser. Qualitative content analysis of programming errors. In *Proceedings of the 5th International Conference on Information and Education Technology*, pages 161–166. ACM, 2017.

[Shanmuganathan 2016] Subana Shanmuganathan. Artificial neural network modelling: An introduction. In *Artificial Neural Network Modelling*, pages 1–14. Springer, 2016.

[Shao *et al.* 2019] Sicong Shao, Cihan Tunc, Amany Al-Shawi, and Salim Hariri. Automated twitter author clustering with unsupervised learning for social media forensics. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE, 2019.

[Sharaff *et al.* 2016] Aakanksha Sharaff, Naresh Kumar Nagwani, and Abhishek Dhadse. Comparative study of classification algorithms for spam email detection. In *Emerging research in computing, information, communication and applications*, pages 237–244. Springer, 2016.

[Shargabi *et al.* 2015] Amal Shargabi, Syed Ahmad Aljunid, Muthukkaruppanan Annamalai, Shuhaida Mohamed Shuhidan, and Abdullah Mohd Zin. Program comprehension levels of abstraction for novices. In *International Conference on Computer, Communications, and Control Technology*, pages 211–215. IEEE, 2015.

[Sharieh *et al.* 2021] Ahmad Sharieh, Ruba Abu Khurmah, Raja Masadeh, Abdullah Alzaqebah, Nesreen Alsharman, and Farah Sharieh. Effect of threat control management strategies on number infected by covid-19. In *The Effect of Coronavirus Disease (COVID-19) on Business Intelligence*, pages 15–41. Springer, 2021.

[Sharma *et al.* 2018] Abhay Sharma, Ananya Nandan, and Reetika Ralhan. An investigation of supervised learning methods for authorship attribution in short hinglish texts using char & word n-grams. *arXiv preprint arXiv:1812.10281*, 2018.

[Sharma *et al.* 2019] Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*, 2019.

[Sharmila and Sakthi 2018] L Sharmila and U Sakthi. Chronological pattern exploration algorithm for gene expression data clustering and classification. *Wireless Personal Communications*, 102(2):1503–1519, 2018.

[Shawe-Taylor *et al.* 2004] John Shawe-Taylor, Nello Cristianini, et al. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[Shekarpour *et al.* 2015] Saeedeh Shekarpour, Edgard Marx, Axel-Cyrille Ngonga Ngomo, and Sören Auer. Sina: Semantic interpretation of user queries for question answering on interlinked data. *Journal of Web Semantics*, 30:39–51, 2015.

[Sheldon 2011] Brian Sheldon. *Cognitive-behavioural therapy: research and practice in health and social care*. Routledge, 2011.

[Shen *et al.* 2016] Zebang Shen, Hui Qian, Tengfei Zhou, and Tongzhou Mu. Adaptive variance reducing for stochastic gradient descent. In *IJCAI*, pages 1990–1996, 2016.

[Shen *et al.* 2017] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in neural information processing systems*, pages 6830–6841, 2017.

[Sheridan 1959] Peter B Sheridan. The arithmetic translator-compiler of the IBM FORTRAN automatic coding system. *Communications of the ACM*, 2(2):9–21, 1959.

[Sherman 1888] Lucius Adelno Sherman. *Some observations upon the sentence-length in English prose*. 1888.

[Shneiderman 1978] Ben Shneiderman. Improving the human factors aspect of database interactions. *ACM Transactions on Database Systems (TODS)*, 3(4):417–439, 1978.

[Shneiderman 2000] Ben Shneiderman. Creating creativity: user interfaces for supporting innovation. *ACM Transactions on Computer-Human Interaction*, 7(1):114–138, 2000.

[Shrestha *et al.* 2017] Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 669–674, 2017.

[Shue *et al.* 2009] Craig A Shue, Minaxi Gupta, John J Lubia, Chin Hua Kong, and Asim Yuksel. Spamology: A study of spam origins. In *the 6th Conference on Email and Anti-Spam (CEAS)*, 2009.

[Shukla 2019] Rohan Shukla. *Foundations and Applications of Authorship Attribution Analysis*. PhD thesis, 2019.

[Shum *et al.* 2018] Heung-Yeung Shum, Xiao-dong He, and Di Li. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, 2018.

[Siau and Tan 2006] Keng Siau and Xin Tan. Cognitive mapping techniques for user-database interaction. *IEEE transactions on professional communication*, 49(2):96–108, 2006.

[Sidorov *et al.* 2012] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic dependency-based n-grams as classification features. In *Mexican International Conference on Artificial Intelligence*, pages 1–11. Springer, 2012.

[Sidorov *et al.* 2013] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic dependency-based n-grams as classification features. In *Advances in Computational Intelligence: 11th Mexican International Conference on Artificial Intelligence, MICAI 2012, San Luis Potosí, Mexico, October 27–November 4, 2012. Revised Selected Papers, Part II 11*, pages 1–11. Springer, 2013.

[Silva *et al.* 2016] Yasin N Silva, Isadora Almeida, and Michell Queiroz. SQL: From traditional databases to big data. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pages 413–418. ACM, 2016.

[Simaki *et al.* 2016] Vasiliki Simaki, Iosif Mporas, and Vasileios Megalooikonomou. Age identification of twitter users: Classification methods and sociolinguistic analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 385–395. Springer, 2016.

[Singh *et al.* 2013] Rishabh Singh, Sumit Gulwani, and Armando Solar-Lezama. Automated feedback generation for introductory programming assignments. *ACM Special Interest Group on Programming Languages*, 48(6):15–26, 2013.

[Singh *et al.* 2021] Rhia Singh, Janith Weerasinghe, and Rachel Greenstadt. Writing style change detection on multi-author documents. In *CLEF (Working Notes)*, pages 2137–2145, 2021.

[Sinha *et al.* 2015] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june Paul Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246. ACM, 2015.

[Sipser 2006] Michael Sipser. *Introduction to the Theory of Computation*, volume 2. Thomson Course Technology Boston, 2006.

[Smith 1983] Michael William Smith. Recent experience and new developments of methods for the determination of authorship. *ALLC BULL.*, 11(3):73–82, 1983.

[Socher *et al.* 2012] Richard Socher, Yoshua Bengio, and Christopher D Manning. Deep learning for NLP (without magic). In *Tutorial Abstracts of ACL 2012*, pages 5–5. Association for Computational Linguistics, 2012.

[Soflano *et al.* 2015] Mario Soflano, Thomas M Connolly, and Thomas Hainey. Learning style analysis in adaptive GBL application to teach SQL. *Computers & Education*, 86:105–119, 2015.

[Sokolov *et al.* 2016] Artem Sokolov, Julia Kreutzer, Christopher Lo, and Stefan Riezler. Learning structured predictors from bandit feedback for interactive NLP. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1610–1620, 2016.

[Sokolova *et al.* 2006] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer, 2006.

[Solorio *et al.* 2011] Thamar Solorio, Sangita Pillay, Sindhu Raghavan, and Manuel Montes y Gómez. Modality specific meta features for authorship attribution in web forum posts. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 156–164, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

[Soloway and Ehrlich 1984] Elliot Soloway and Kate Ehrlich. Empirical studies of programming knowledge. *IEEE Transactions on software engineering*, (5):595–609, 1984.

[Song *et al.* 2011] Qinbao Song, Jingjie Ni, and Guangtao Wang. A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE transactions on knowledge and data engineering*, 25(1):1–14, 2011.

[Song *et al.* 2015] Dezhao Song, Frank Schilder, Charese Smiley, Chris Brew, Tom Zielund, Hiroko Bretz, Robert Martin, Chris Dale, John Duprey, Tim Miller, et al. Tr discover: A natural language interface for querying and analyzing interlinked datasets. In *International Semantic Web Conference*, pages 21–37. Springer, 2015.

[Sonnhammer *et al.* 1998] Erik LL Sonnhammer, Gunnar Von Heijne, Anders Krogh, et al. A hidden Markov model for predicting transmembrane helices in protein sequences. In *Ismb*, volume 6, pages 175–182, 1998.

[Sowell *et al.* 2009] Benjamin Sowell, Alan Demers, Johannes Gehrke, Nitin Gupta, Haoyuan Li, and Walker White. From declarative languages to declarative processing in computer games. *arXiv preprint arXiv:0909.1770*, 2009.

[Soylu *et al.* 2016] Ahmet Soylu, Martin Giese, Ernesto Jimenez-Ruiz, Guillermo Vega-Gorgojo, and Ian Horrocks. Experiencing optiqueVQS: a multi-paradigm and ontology-based visual query system for end users. *Universal Access in the Information Society*, 15(1):129–152, 2016.

[Soylu *et al.* 2017] Ahmet Soylu, Martin Giese, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Dmitriy Zheleznyakov, and Ian Horrocks. Ontology-based end-user visual query formulation: Why, what, who, how, and which? *Universal Access in the Information Society*, 16(2):435–467, 2017.

[Springer 1923] J. F. Springer. A Mechanical Solution of a Literary Problem. *Scientific American*, 128(2):100–100, February 1923.

[Srinivasa-Desikan 2018] Bhargav Srinivasa-Desikan. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd, 2018.

[Srinivasan and Brown 2002] Savitha Srinivasan and Eric Brown. Is speech recognition becoming mainstream? *Computer*, (4):38–41, 2002.

[Srivastava *et al.* 2015] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385, 2015.

[Stamatatos and Koppel 2011] Efstathios Stamatatos and Moshe Koppel. Plagiarism and authorship analysis: introduction to the special issue. *Language Resources and Evaluation*, 45(1):1–4, 2011.

[Stamatatos and others 2006] Efstathios Stamatatos et al. Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, volume 36, pages 41–46, 2006.

[Stamatatos *et al.* 2000] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495, 2000.

[Stamatatos *et al.* 2001] Efstathios Stamatatos, Nikos Fakotakis, and Georgios Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[Stamatatos *et al.* 2014] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A Sanchez-Perez, and Alberto Barrón-Cedeño. Overview of the author identification task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–21, 2014.

[Stamatatos *et al.* 2022] Efstathios Stamatatos, Mike Kestemont, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast. Overview of the authorship verification task at pan 2022. In *CEUR workshop proceedings*, volume 3180, pages 2301–2313, 2022.

[Stamatatos 2007] Efstathios Stamatatos. Author identification using imbalanced and limited training texts. In *18th International Workshop on Database and Expert Systems Applications (DEXA 2007)*, pages 237–241. IEEE, 2007.

[Stamatatos 2008] Efstathios Stamatatos. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, 44(2):790–799, 2008.

[Stamatatos 2009a] Efstathios Stamatatos. Intrinsic plagiarism detection using character n-gram profiles. *threshold*, 2(1,500), 2009.

[Stamatatos 2009b] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

[Stamatatos 2013] Efstathios Stamatatos. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21(2):421–439, 2013.

[Stamatatos 2018] Efstathios Stamatatos. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and technology*, 69(3):461–473, 2018.

[Stańczyk and Cyran 2007] Urszula Stańczyk and Krzysztof A Cyran. Machine learning approach to authorship attribution of literary texts. *International journal of applied mathematics and informatics*, 1(4):151–158, 2007.

[Starr *et al.* 2008] Christopher W Starr, Bill Manaris, and RoxAnn H Stalvey. Bloom's taxonomy revisited: specifying assessable learning objectives in computer science. In *ACM SIGCSE Bulletin*, volume 40, pages 261–265. ACM, 2008.

[Statista] Statista. *Annual Twitter user growth rate in the United States from 2020 to 2025*. https://www.statista.com/statistics/238729/twitters-annual-growth-rate-in-the-us/. Accessed: 2022-09-30.

[Staykova 2014] Kamenka Staykova. Natural language generation and semantic technologies. *Cybernetics and Information Technologies*, 14(2):3–23, 2014.

[Stein *et al.* 2007] Benno Stein, Moshe Koppel, and Efstathios Stamatatos. Plagiarism analysis, authorship identification, and near-duplicate detection (pan'07). In *SIGIR Forum*, volume 41, pages 68–71, 2007.

[Stein *et al.* 2011] Benno Stein, Nedim Lipka, and Peter Prettenhofer. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82, 2011.

[Stein *et al.* 2019] Roger Alan Stein, Patricia A Jaques, and João Francisco Valiati. An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471:216–232, 2019.

[Stokel-Walker 2023] Chris Stokel-Walker. *Why is Twitter becoming X?*, 2023.

[Stolerman *et al.* 2014] Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. Breaking the closed-world assumption in stylometric authorship attribution. In *IFIP International Conference on Digital Forensics*, pages 185–205. Springer, 2014.

[Storey 2005] M-A Storey. Theories, methods and tools in program comprehension: Past, present and future. In *13th International Workshop on Program Comprehension*, pages 181–191. IEEE, 2005.

[Storey 2006] Margaret-Anne Storey. Theories, tools and research methods in program comprehension: past, present and future. *Software Quality Journal*, 14(3):187–208, 2006.

[Straka and Straková 2017] Milan Straka and Jana Straková. Tokenizing, PoS tagging, lemmatizing and parsing UD 2.0 with UDpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, 2017.

[Štrumbelj and Kononenko 2014] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

[Sturmfels *et al.* 2018] Pascal Sturmfels, Saige Rutherford, Mike Angstadt, Mark Peterson, Chandra Sripada, and Jenna Wiens. A domain guided CNN architecture for predicting age from structural brain images. *arXiv preprint arXiv:1808.04362*, 2018.

[Sujatha *et al.* 2012] B Sujatha, Dr S Viswanadha Raju, and Humera Shaziya. A survey of natural language interface to database management system. *International Journal of Science and Advance Technology*, 2(6):56–61, 2012.

[Sultana and Gavrilova 2018] Madeena Sultana and Marina Gavrilova. Temporal Pattern in Tweeting Behavior for Persons' Identity Verification. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2472–2477. IEEE, 2018.

[Sultana *et al.* 2014] Madeena Sultana, Padma Polash Paul, and Marina Gavrilova. Mining social behavioral biometrics in Twitter. In *2014 International Conference on Cyberworlds*, pages 293–299. IEEE, 2014.

[Sultana *et al.* 2017] Madeena Sultana, Padma Polash Paul, and Marina L Gavrilova. User recognition from social behavior in computer-mediated social context. *IEEE Transactions on Human-Machine Systems*, 47(3):356–367, 2017.

[Suman *et al.* 2021] Chanchal Suman, Ayush Raj, Sriparna Saha, and Pushpak Bhattacharyya. Authorship attribution of microtext using capsule networks. *IEEE Transactions on Computational Social Systems*, 2021.

[Sun *et al.* 2017a] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[Sun *et al.* 2017b] Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. *Information fusion*, 36:10–25, 2017.

[Sundararajan and Woodard 2018] Kalaivani Sundararajan and Damon Woodard. What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822, 2018.

[Sung *et al.* 2013] Juyup Sung, Seunghyeon Moon, and Jae-Gil Lee. The influence in twitter: Are they really influenced? In Longbing Cao, Hiroshi Motoda, Jaideep Srivastava, Ee-Peng Lim, Irwin King, Philip S. Yu, Wolfgang Nejdl, Guandong Xu, Gang Li, and Ya Zhang, editors, *Behavior and Social Computing*, pages 95–105, Cham, 2013. Springer International Publishing.

[Suranauwarat 2017] Sukanya Suranauwarat. An approach to solving technical difficulties facing non-CS students in a database class. *International Journal of Modern Education and Computer Science*, 9(2):14, 2017.

[Sutton and Staw 1995] Robert I. Sutton and Barry M. Staw. What theory is not. *Administrative Science Quarterly*, 40(3):371–384, 1995.

[Suzuki and Nagata 2015] Jun Suzuki and Masaaki Nagata. A unified learning framework of skip-grams and global vectors. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 186–191, 2015.

[Swain *et al.* 2017] Siddharth Swain, Gaurav Mishra, and C Sindhu. Recent approaches on authorship attribution techniques—an overview. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 1, pages 557–566. IEEE, 2017.

[Sweller 1988] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.

[Sweller 2006] John Sweller. The worked example effect and human cognition. *Learning and instruction*, 16(2):165–169, 2006.

[Sweta and Lal 2016] Soni Sweta and Kanhaiya Lal. Learner model for automatic detection of learning style using FCM in adaptive e-learning system. *IOSR Journal of Engineering*, 18(2):18–24, 2016.

[Taha *et al.* 2021] Ahmed Taha, Heba M Khalil, and Tarek El-Shishtawy. A two level learning model for authorship authentication. *Plos one*, 16(8):e0255661, 2021.

[Taipalus *et al.* 2018] Toni Taipalus, Mikko Siponen, and Tero Vartiainen. Errors and complications in SQL query formulation. *ACM Transactions on Computing Education (TOCE)*, 18(3):15, 2018.

[Taira and Haruno 1999] Hirotoshi Taira and Masahiko Haruno. Feature selection in svm text categorization. In *AAAI/IAAI*, pages 480–486, 1999.

[Tan and Tsai 2010] Richmond Hong Rui Tan and Flora S. Tsai. Authorship identification for online text. In *Proceedings of the 2010 International Conference on Cyberworlds*, CW '10, page 155â€"162, USA, 2010. IEEE Computer Society.

[Tanaka and Jin 2014] Ryoko Tanaka and Mingzhe Jin. Authorship attribution of cell-phone e-mail. *International Information Institute (Tokyo). Information*, 17(4):1217, 2014.

[Tang *et al.* 2014] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37, 2014.

[Tang *et al.* 2016] Bo Tang, Haibo He, Paul M Baggenstoss, and Steven Kay. A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1602–1606, 2016.

[Tang 2013] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[Teh *et al.* 2004] Yee Teh, Michael Jordan, Matthew Beal, and David Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, 17, 2004.

[Thelwall *et al.* 2012] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[Theophilo *et al.* 2021] Antonio Theophilo, Romain Giot, and Anderson Rocha. Authorship attribution of social media messages. *IEEE Transactions on Computational Social Systems*, 2021.

[Thomas 2014] Jenny A Thomas. *Meaning in interaction: An introduction to pragmatics*. Routledge, 2014.

[Thompson *et al.* 2008] Errol Thompson, Andrew Luxton-Reilly, Jacqueline L Whalley, Minjie Hu, and Phil Robbins. Bloom's taxonomy for CS assessment. In *Proceedings of the tenth conference on Australasian computing education-Volume 78*, pages 155–161. Australian Computer Society, Inc., 2008.

[Thurley and Dennick 2008] P Thurley and R Dennick. Problem-based learning and radiology. *Clinical radiology*, 63(6):623–628, 2008.

[Tidwell 1999] Jenifer Tidwell. *Common ground: A pattern language for human-computer interface design*, 1999.

[Tidwell 2002] Jenifer Tidwell. UI patterns and techniques. *Zdroj: www. mit. edu/˜ jtidwell*, 2002.

[Tidwell 2010] Jenifer Tidwell. *Designing interfaces: Patterns for effective interaction design*. " O'Reilly Media, Inc.", 2010.

[Tittizer and Ramamoorthy 2006] Abigail Tittizer and Venkatesan Ramamoorthy. *Systems and methods for managing data associated with computer code*, 2006. US Patent App. 11/379,390.

[Tobin 2012] Kenneth G Tobin. Constructivism as a referent for teaching and learning. In *The practice of constructivism in science education*, pages 19–38. Routledge, 2012.

[Tobon-Mejia *et al.* 2011] Diego A Tobon-Mejia, Kamal Medjaher, Noureddine Zerhouni, and Gérard Tripot. Hidden Markov models for failure diagnostic and prognostic. In *2011 Prognostics and System Health Managment Confernece*, pages 1–8. IEEE, 2011.

[Toraman *et al.* 2022] Cagri Toraman, Furkan Şahinuç, Eyup Halit Yilmaz, and Ibrahim Batuhan Akkaya. Understanding social engagements: A comparative analysis of user and text features in twitter. *Social network analysis and mining*, 12(1):47, 2022.

[Torney *et al.* 2012] Rosemary Torney, Peter Vamplew, and John Yearwood. Using psycholinguistic features for profiling first language of authors. *Journal of the American Society for Information Science and Technology*, 63(6):1256–1269, 2012.

[Trask 2004] Robert Lawrence Trask. *A dictionary of phonetics and phonology*. Routledge, 2004.

[Tratz *et al.* 2007] Stephen Tratz, Antonio Sanfilippo, Michelle Gregory, Alan Chappell, Christian Posse, and Paul Whitney. Pnnl: A supervised maximum entropy approach to word sense disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 264–267. Association for Computational Linguistics, 2007.

[Treiber *et al.* 2011] Martin Treiber, Daniel Schall, Schahram Dustdar, and Christian Scherling. Tweetflows: flexible workflows with twitter. In *Proceedings of the 3rd international workshop on Principles of engineering service-oriented systems*, pages 1–7, 2011.

[Truong 2016] Huong May Truong. Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities. *Computers in Human Behavior*, 55:1185–1193, 2016.

[Tsarfaty *et al.* 2010] Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Marie Candito, Jennifer Foster, Yannick Versley, Ines Rehbein, and Lamia Tounsi. Statistical parsing of morphologically rich languages (SPMRL): what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 1–12. Association for Computational Linguistics, 2010.

[Tschuggnall *et al.* 2017] Michael Tschuggnall, Efstthios Stamatatos, Ben Verhoeven, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. Overview of the author identification task at pan-2017: style breach detection and author clustering. In *Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al.*, pages 1–22, 2017.

[Tsujii 2000] Junâ€™ichi Tsujii. Invited talk: Generic nlp technologies: Language, knowledge and information extraction. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 12–22, 2000.

[Tufekci 2014] Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 505–514, 2014.

[Tukey and others 1977] John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, MA, 1977.

[Tukey 1980] John W Tukey. We need both exploratory and confirmatory. *The american statistician*, 34(1):23–25, 1980.

[Turchin *et al.* 2006] Alexander Turchin, Nikheel S Kolatkar, Richard W Grant, Eric C Makhni, Merri L Pendergrass, and Jonathan S Einbinder. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association*, 13(6):691–695, 2006.

[Turell and Gibbons 2008] M Teresa Turell and John Gibbons. Dimensions of forensic linguistics. *Dimensions of Forensic Linguistics*, pages 1–324, 2008.

[Twain 2013] Mark Twain. Morphology: the words of language. 2013.

[Twitter, Inc 2022] Twitter, Inc. *Twitter Announces Second Quarter 2022 Results*. https://investor.twitterinc.com/financial-information/quarterly-results/default.aspx, 2022. [Online; accessed 11-September-2022].

[Twitter 2022] Twitter. *About Twitter | Our company purpose, principles, leadership*. https://about.twitter.com/en/who-we-are/our-company, 2022. [Online; accessed 05-May-2022].

[Tyo *et al.* 2022] Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. *On the State of the Art in Authorship Attribution and Authorship Verification*, 2022.

[Tzotsos and Argialas 2008] Angelos Tzotsos and Demetre Argialas. Support vector machine classification for object-based image analysis. *Object-based image analysis: Spatial concepts for knowledge-driven remote sensing applications*, pages 663–677, 2008.

[UÄŸuz 2011] Harun UÄŸuz. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*, 24(7):1024–1032, 2011.

[Uchendu *et al.* 2023] Adaku Uchendu, Thai Le, and Dongwon Lee. Attribution and obfuscation of neural text authorship: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 25(1):1–18, 2023.

[Ullman and Aho 1977] Jeffrey D Ullman and Alfred V Aho. Principles of compiler design. *Reading: Addison Wesley*, 1977.

[Urribarri and Castro 2017] Dana K Urribarri and Silvia M Castro. Prediction of data visibility in two-dimensional scatterplots. *Information Visualization*, 16(2):113–125, 2017.

[Urvoy *et al.* 2008] Tanguy Urvoy, Emmanuel Chauveau, Pascal Filoche, and Thomas Lavergne. Tracking web spam with html style similarities. *ACM Transactions on the Web (TWEB)*, 2(1):1–28, 2008.

[Utama *et al.* 2017] Prasetya Utama, Nathaniel Weir, Carsten Binnig, and Ugur Çetintemel. Voice-based data exploration: Chatting with your database. In *Proceedings of the 2017 workshop on Search-Oriented Conversational AI*, pages 1–6, 2017.

[Utama *et al.* 2018] Prasetya Utama, Nathaniel Weir, Fuat Basik, Carsten Binnig, Ugur Cetintemel, Benjamin Hättasch, Amir Ilkhechi, Shekar Ramaswamy, and Arif Usta. An end-to-end neural natural language interface for databases. *arXiv preprint arXiv:1804.00401*, 2018.

[Valli and Lucas 2000] Clayton Valli and Ceil Lucas. *Linguistics of American sign language: An introduction*. Gallaudet University Press, 2000.

[Van den Brand *et al.* 2005] Mark Van den Brand, P-E Moreau, and Jurgen Vinju. Generator of efficient strongly typed abstract syntax trees in Java. *IEE Proceedings-Software*, 152(2):70–78, 2005.

[Van der Maaten and Hinton 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Van Gog *et al.* 2011] Tamara Van Gog, Liesbeth Kester, and Fred Paas. Effects of worked examples, example-problem, and problem-example pairs on novicesâ€™ learning. *Contemporary Educational Psychology*, 36(3):212–218, 2011.

[Van Halteren *et al.* 2005] Hans Van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77, 2005.

[Van Kasteren *et al.* 2010] TLM Van Kasteren, Gwenn Englebienne, and Ben JA Kröse. Activity recognition using semi-Markov models on real world smart home datasets. *Journal of ambient intelligence and smart environments*, 2(3):311–325, 2010.

[Van Wanrooij and Pras 2010] Ward Van Wanrooij and Aiko Pras. Filtering spam from bad neighborhoods. *International Journal of Network Management*, 20(6):433–444, 2010.

[Van Welie and Van der Veer 2003] Martijn Van Welie and Gerrit C Van der Veer. Pattern languages in interaction design: Structure and organization. In *Proceedings of interact*, volume 3, pages 1–5, 2003.

[Vapnik and Chervonenkis 1964] V Vapnik and A Ya Chervonenkis. A class of algorithms for pattern recognition learning. *Avtomat. i Telemekh*, 25(6):937–945, 1964.

[Vaswani *et al.* 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Venčkauskas *et al.* 2015] Algimantas Venčkauskas, Robertas Damaševičius, Romas Marcinkevičius, and Arnas Karpavičius. Problems of authorship identification of the national language electronic discourse. In *International Conference on Information and Software Technologies*, pages 415–432. Springer, 2015.

[Verduyn *et al.* 2017] Philippe Verduyn, Oscar Ybarra, Maxime Résibois, John Jonides, and Ethan Kross. Do social network sites enhance or undermine subjective well-being? a critical review. *Social Issues and Policy Review*, 11(1):274–302, 2017.

[Verduyn *et al.* 2020] Philippe Verduyn, Nino Gugushvili, Karlijn Massar, Karin Täht, and Ethan Kross. Social comparison on social networking sites. *Current opinion in psychology*, 36:32–37, 2020.

[Verenna *et al.* 2018] Anne-Marie A Verenna, Kim A Noble, Helen E Pearson, and Susan M Miller. Role of comprehension on performance at higher levels of Bloom's taxonomy: Findings from assessments of healthcare professional students. *Anatomical sciences education*, 11(5):433–444, 2018.

[Vergara and Estévez 2014] Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1):175–186, 2014.

[Verma and Srinivasan 2019a] Gaurav Verma and Balaji Vasan Srinivasan. A lexical, syntactic, and semantic perspective for understanding style in text. *arXiv preprint arXiv:1909.08349*, 2019.

[Verma and Srinivasan 2019b] Gaurav Verma and Balaji Vasan Srinivasan. *A Lexical, Syntactic, and Semantic Perspective for Understanding Style in Text*. https://arxiv.org/abs/1909.08349, 2019.

[Verma *et al.* 2019] Amit Verma, Kirill M Yurov, Peggy L Lane, and Yuliya V Yurova. An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements. *Journal of Education for Business*, 94(4):243–250, 2019.

[Vermunt and Donche 2017] Jan D Vermunt and Vincent Donche. A learning patterns perspective on student learning in higher education: state of the art and moving forward. *Educational Psychology Review*, 29(2):269–299, 2017.

[Vilar and Federico 2021] David Vilar and Marcello Federico. A statistical extension of byte-pair encoding. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 263–275, 2021.

[Von Mayrhauser and Vans 1993] Anneliese Von Mayrhauser and A Marie Vans. From program comprehension to tool requirements for an industrial environment. In *[1993] IEEE Second Workshop on Program Comprehension*, pages 78–86. IEEE, 1993.

[Von Mayrhauser and Vans 1995] Anneliese Von Mayrhauser and A Marie Vans. Program comprehension during software maintenance and evolution. *Computer*, 28(8):44–55, 1995.

[Voorhees and Harman 1999] Ellen M Voorhees and Donna Harman. The Text REtrieval Conference (trec): History and plans for TREC-9. In *ACM SIGIR Forum*, volume 33, pages 12–15. ACM, 1999.

[Vosoughi *et al.* 2016] Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1041–1044, 2016.

[Voulodimos *et al.* 2018] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[Vu *et al.* 2018] Thanh Vu, Dat Quoc Nguyen, Xuan-Son Vu, Michael Catt, Michael Trenell, et al. Nihrio at semeval-2018 task 3: A simple and accurate neural network model for irony detection in twitter. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 525–530, 2018.

[Vukovic 2007] Nina Tudman Vukovic. David crystal: Language and the internet. *Suvremena Lingvistika*, 33(63):135–136, 2007.

[Wadhwa and Palvia 2018] Vivek Wadhwa and Shailendra Palvia. Is information technology hacking our happiness? *Journal of Information Technology Case and Application Research*, 20(3-4):151–157, 2018.

[Walkley 2010] Amelia Walkley. Typing friendship into being: vocatives in facebook wall-to-wall conversations. 2010.

[Wallace 2009] Richard S Wallace. The anatomy of ALICE. In *Parsing the Turing Test*, pages 181–210. Springer, 2009.

[Walter *et al.* 2012] Sebastian Walter, Christina Unger, Philipp Cimiano, and Daniel Bär. Evaluation of a layered approach to question answering over linked data. In *International Semantic Web Conference*, pages 362–374. Springer, 2012.

[Wang and Zong 2010] Zhiguo Wang and Chengqing Zong. Phrase structure parsing with dependency structure. In *Coling 2010: Posters*, pages 1292–1300, 2010.

[Wang *et al.* 2006] Kevin Wang, Corey McCaffrey, Daniel Wendel, and Eric Klopfer. 3D game design with programming blocks in StarLogo TNG. In *Proceedings of the 7th international conference on Learning sciences*, pages 1008–1009. International Society of the Learning Sciences, 2006.

[Wang *et al.* 2015] Fei Wang, Cai-hong Li, Jing-shan Wang, Jiao Xu, and Lian Li. A two-stage feature selection method for text categorization by using category correlation degree and latent semantic indexing. *Journal of Shanghai Jiaotong University (Science)*, 20:44–50, 2015.

[Wang *et al.* 2016a] Jin Wang, Liang-Chih Yu, K Robert Lai, and Xuejie Zhang. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 225–230, 2016.

[Wang *et al.* 2016b] Xinggang Wang, Buyun Sheng, Liang Xue, and Zheng Xiao. Classification of customer requirements on map reduce-based naive Bayes. In *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, pages 1–4. IEEE, 2016.

[Wang *et al.* 2017a] Chenglong Wang, Alvin Cheung, and Rastislav Bodik. Interactive query synthesis from input-output examples. In *Proceedings of the ACM International Conference on Management of Data*, pages 1631–1634. ACM, 2017.

[Wang *et al.* 2017b] Chenglong Wang, Alvin Cheung, and Rastislav Bodik. Synthesizing highly expressive SQL queries from input-output examples. In *ACM Special Interest Group on Programming Languages Notices*, volume 52, pages 452–466. ACM, 2017.

[Wang *et al.* 2017c] Qi Wang, Jungang Xu, Hong Chen, and Ben He. Two improved continuous bag-of-word models. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2851–2856, May 2017.

[Wang *et al.* 2017d] Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, Neil O'Hare, and Yi Chang. Understanding and discovering deliberate self-harm content in social media. In *Proceedings of the 26th international conference on World Wide Web*, pages 93–102, 2017.

[Wang *et al.* 2019] Yang Wang, Paulo Goes, Zaiyan Wei, and Daniel Zeng. Production of online word-of-mouth: Peer effects and the moderation of user characteristics. *Production and Operations Management*, 28(7):1621–1640, 2019.

[Wang 2010] Alex Hai Wang. Don't follow me: Spam detection in twitter. In *2010 International Conference on Security and Cryptography (SECRYPT)*, pages 1–10, July 2010.

[Wängqvist and Frisén 2016] Maria Wängqvist and Ann Frisén. Who am i online? understanding the meaning of online contexts for identity development. *Adolescent Research Review*, 1(2):139–151, 2016.

[Wanner and others 2017] Leo Wanner et al. On the relevance of syntactic and discourse features for author profiling and identification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 681–687, 2017.

[Ward 2015] Patrick T Ward. *Teaching of SQL Through a Game*. PhD thesis, 2015.

[Ward 2016] Gregory L Ward. *The semantics and pragmatics of preposing*. Routledge, 2016.

[Warner and Hirschberg 2012] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics, 2012.

[Warnke 2009] Elizabeth Warnke. Technical writing for software documentation writers: A textbook on process and product. 2009.

[Warren and Pereira 1982] D Warren and F Pereira. An efficient for interpreting easily adaptable system natural language queries. *American Journal*, 8(3-4):110–122, 1982.

[Warth *et al.* 2008] Alessandro Warth, Takashi Yamamiya, Yoshiki Ohshima, and Scott Wallace. Toward a more scalable end-user scripting language. In *Sixth International Conference on Creating, Connecting and Collaborating through Computing*, pages 172–178. IEEE, 2008.

[Wasserman and Faust 1994] Stanley Wasserman and Katherine Faust. Social network analysis: Methods and applications. 1994.

[Watson *et al.* 2011] Christopher Watson, Frederick WB Li, and Rynson WH Lau. Learning programming languages through corrective feedback and concept visualisation. In *International Conference on Web-Based Learning*, pages 11–20. Springer, 2011.

[Weizenbaum and others 1966] Joseph Weizenbaum et al. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[Weller *et al.* 2014] Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann. Twitter and society: An introduction. *Twitter and society [Digital Formations, Volume 89]*, pages xxix–xxxviii, 2014.

[Wen and Li 2018] Shuang Wen and Jian Li. Recurrent convolutional neural network with attention for twitter and yelp sentiment classification: Arc model for sentiment classification. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–7, 2018.

[Weren *et al.* 2014] Edson RD Weren, Anderson U Kauer, Lucas Mizusaki, Viviane P Moreira, J Palazzo M de Oliveira, and Leandro K Wives. Examining multiple features for author profiling. *Journal of information and data management*, 5(3):266–266, 2014.

[Weston *et al.* 2012] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012.

[Whiting and Williams 2013] Anita Whiting and David Williams. Why people use social media: a uses and gratifications approach. *Qualitative market research: an international journal*, 16(4):362–369, 2013.

[Whittaker and Kowalski 2015] Elizabeth Whittaker and Robin M Kowalski. Cyberbullying via social media. *Journal of school violence*, 14(1):11–29, 2015.

[Whittaker 2010] L Whittaker. Facebook speak: Teenagers create secret online language. *The Telegraph*, 2010.

[Wiebe *et al.* 2005] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2):165–210, 2005.

[Wiedenbeck 1986] Susan Wiedenbeck. Beacons in computer program comprehension. *International Journal of Man-Machine Studies*, 25(6):697–709, 1986.

[Wijngaards *et al.* 2022] Indy Wijngaards, Florie R. Pronk, and Martijn J. Burger. For whom and under what circumstances does email message batching work? *Internet Interventions*, 27:100494, 2022.

[Wikipedia contributors 2022a] Wikipedia contributors. *Jaiku — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Jaiku&oldid=1121928249, 2022. [Online; accessed 19-January-2023].

[Wikipedia contributors 2022b] Wikipedia contributors. *Twitter — Wikipedia, The Free Encyclopedia*, 2022. [Online; accessed 19-August-2022].

[Wikipedia contributors 2023a] Wikipedia contributors. *Hacktivism — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Hacktivism&oldid=1140385092, 2023. [Online; accessed 1-May-2023].

[Wikipedia contributors 2023b] Wikipedia contributors. *The Shadow Brokers — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=The_Shadow_Brokers&oldid=1151688879, 2023. [Online; accessed 1-May-2023].

[Wikipedia 2022] Wikipedia. *Twitter — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Twitter&oldid=1109006371, 2022. [Online; accessed 11-September-2022].

[Wikisource 2018] Wikisource. *Popular Science Monthly/Volume 60/December 1901/A Mechanical Solution of a Literary Problem — Wikisource,*, 2018. [Online; accessed 19-October-2022].

[Wilcock 2009] Graham Wilcock. Text annotation with openNLP and Uima. 2009.

[Wilcocks and Sanders 1994] Derek Wilcocks and Ian Sanders. Animating recursion as an aid to instruction. *Computers & Education*, 23(3):221–226, 1994.

[Wilhelm *et al.* 2013] Reinhard Wilhelm, Helmut Seidl, and Sebastian Hack. *Compiler design: Syntactic and semantic analysis*. Springer Science & Business Media, 2013.

[Williams 1975a] Carrington B Williams. Mendenhall's studies of word-length distribution in the works of shakespeare and bacon. *Biometrika*, 62(1):207–212, 1975.

[Williams 1975b] Carrington B Williams. Mendenhall's studies of word-length distribution in the works of shakespeare and bacon. *Biometrika*, 62(1):207–212, 1975.

[Wilson *et al.* 2010] Dale-Marie Wilson, Aqueasha M Martin, and Juan E Gilbert. How may i help you'-spoken queries for technical assistance. In *Proceedings of the 48th Annual Southeast Regional Conference*, page 43. ACM, 2010.

[Wilson 2018] Leslie Owen Wilson. Anderson and krathwohl–Bloomâ€™s taxonomy revised. *Accessed online: https://thesecondprinciple. com/teaching-essentials/beyond-bloom-cognitive-taxonomy-revised*, 2018.

[Winograd 1973] Terry Winograd. A procedural model of language understanding. 1973.

[Winthrop and McGivney 2015] Rebecca Winthrop and Eileen McGivney. Why wait 100 years? bridging the gap in global education. *The Brookings Institution*, 2015.

[Wolfe and Kolb 1984] Donald M Wolfe and David A Kolb. Career development, personal growth and experiential learning. *D. Kolb, IM Rubin and JM McIntyre op. cit*, 1984.

[Wong *et al.* 2013] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets and retweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 640–649, 2013.

[Wood *et al.* 2019] Alexander Wood, Vladimir Shpilrain, Kayvan Najarian, and Delaram Kahrobaei. Private naive Bayes classification of personal biomedical data: Application in cancer data analysis. *Computers in biology and medicine*, 105:144–150, 2019.

[Woodhams *et al.* 2007] Jessica Woodhams, Tim D Grant, and Andrew RG Price. From marine ecology to crime analysis: Improving the detection of serial sexual offences using a taxonomic similarity measure. *Journal of Investigative Psychology and Offender Profiling*, 4(1):17–27, 2007.

[Woods 1972] William Woods. The lunar sciences natural language information system. *BBN report*, 1972.

[Woods 1973] William A Woods. Progress in natural language understanding: an application to lunar geology. In *Proceedings of national computer conference and exposition*, pages 441–450. ACM, 1973.

[Woods 1978] William A Woods. Semantics and quantification in natural language question answering. In *Advances in computers*, volume 17, pages 1–87. Elsevier, 1978.

[Wu *et al.* 2021a] Haiyan Wu, Zhiqiang Zhang, and Qingfeng Wu. Exploring syntactic and semantic features for authorship attribution. *Applied Soft Computing*, 111:107815, 2021.

[Wu *et al.* 2021b] Haiyan Wu, Zhiqiang Zhang, and Qingfeng Wu. Exploring syntactic and semantic features for authorship attribution. *Applied Soft Computing*, 111:107815, 2021.

[Wudaru *et al.* 2019] Vishal Wudaru, Nikhil Koditala, Aruneswara Reddy, and Radhika Mamidi. Question answering on structured data using NLIDB approach. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pages 1–4. IEEE, 2019.

[Xie *et al.* 2008] Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, and Ivan Osipkov. Spamming botnets: signatures and characteristics. *ACM SIGCOMM Computer Communication Review*, 38(4):171–182, 2008.

[Xu *et al.* 2012] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666, 2012.

[Xu *et al.* 2018] Kun Xu, Lingfei Wu, Zhiguo Wang, Mo Yu, Liwei Chen, and Vadim Sheinin. SQL-to-text generation with graph-to-sequence model. *arXiv preprint arXiv:1809.05255*, 2018.

[Xu 2018] Shuo Xu. Bayesian naïve Bayes classifiers to text classification. *Journal of Information Science*, 44(1):48–59, 2018.

[Yaghmazadeh *et al.* 2017a] Navid Yaghmazadeh, Yuepeng Wang, Isil Dillig, and Thomas Dillig. SQLizer: query synthesis from natural language. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA):63, 2017.

[Yaghmazadeh *et al.* 2017b] Navid Yaghmazadeh, Yuepeng Wang, Isil Dillig, and Thomas Dillig. Type-and content-driven synthesis of SQL queries from natural language. *arXiv preprint arXiv:1702.01168*, 2017.

[Yang and Chow 2014a] Min Yang and Kam-Pui Chow. Authorship attribution for forensic investigation with thousands of authors. In *IFIP International Information Security Conference*, pages 339–350. Springer, 2014.

[Yang and Chow 2014b] Min Yang and Kam-Pui Chow. Authorship attribution for forensic investigation with thousands of authors. In Nora Cuppens-Boulahia, Frédéric Cuppens, Sushil Jajodia, Anas Abou El Kalam, and Thierry Sans, editors, *ICT Systems Security and Privacy Protection*, pages 339–350, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[Yang and Pedersen 1997] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, page 412â€"420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[Yang *et al.* 2012] Jieming Yang, Yuanning Liu, Xiaodong Zhu, Zhen Liu, and Xiaoxu Zhang. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. *Information Processing & Management*, 48(4):741–754, 2012.

[Yang *et al.* 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[Yao *et al.* 2019] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019.

[Yassine *et al.* 2017] Alaeeddine Yassine, Driss Chenouni, Mohammed Berrada, and Ahmed Tahiri. A serious game for learning C programming language concepts using Solo taxonomy. *International Journal of Emerging Technologies in Learning (iJET)*, 12(03):110–127, 2017.

[Yathongchai *et al.* 2017] Wilairat Yathongchai, Jitimon Angskun, and Chun Che FUNG. An ontology model for developing a SQL personalized intelligent tutoring system. *Naresuan University Journal: Science and Technology*, 25(4):88–96, 2017.

[Yu and Ginsberg 2018] Cui Yu and Avi Ginsberg. Social networking without sacrificing privacy. In *2018 International Conference on System Science and Engineering (ICSSE)*, pages 1–6. IEEE, 2018.

[Yu 2012] Sheng Yu. Regular languages. *Handbook of Formal Languages: Volume 1 Word, Language, Grammar*, page 41, 2012.

[Yuan *et al.* 2019] Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4):294–305, 2019.

[Yule 1939] G Udny Yule. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4):363–390, 1939.

[Yun *et al.* 2020] Joseph T Yun, Brittany RL Duff, Patrick T Vargas, Hari Sundaram, and Itai Himelboim. Computationally analyzing social media text for topics: A primer for advertising researchers. *Journal of Interactive Advertising*, 20(1):47–59, 2020.

[Zafarani *et al.* 2019] Reza Zafarani, Xinyi Zhou, Kai Shu, and Huan Liu. Fake news research: Theories, detection strategies, and open problems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 3207â€"3208, New York, NY, USA, 2019. Association for Computing Machinery.

[Zaman *et al.* 2019] Anis Zaman, Rupam Acharyya, Henry Kautz, and Vincent Silenzio. Detecting low self-esteem in youths from web search data. In *The World Wide Web Conference*, pages 2270–2280, 2019.

[Zamfir *et al.* 2019] Vlad-Andrei Zamfir, Mihai Carabas, Costin Carabas, and Nicolae Tapus. Systems monitoring and big data analysis using the elasticsearch system. In *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*, pages 188–193. IEEE, 2019.

[Zangerle *et al.* 2019] Eva Zangerle, Michael Tschuggnall, Günther Specht, M Potthast, and B Stein. Overview of the style change detection task at pan 2019. In *CLEF*, 2019.

[Zechner *et al.* 2009] Mario Zechner, Markus Muhr, Roman Kern, and Michael Granitzer. External and intrinsic plagiarism detection using vector space models. In *Proc. SEPLN*, volume 32, pages 47–55, 2009.

[Zhang and Sun 2013] Sai Zhang and Yuyin Sun. Automatically synthesizing SQL queries from input-output examples. In *2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 224–234. IEEE, 2013.

[Zhang *et al.* 2008] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 21(8):879–886, 2008.

[Zhang *et al.* 2011] Wen Zhang, Taketoshi Yoshida, and Xijin Tang. A comparative study of tf* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765, 2011.

[Zhang *et al.* 2015] Dell Zhang, Jun Wang, and Xiaoxue Zhao. Estimating the uncertainty of average f1 scores. In *Proceedings of the 2015 International conference on the theory of information retrieval*, pages 317–320, 2015.

[Zhang *et al.* 2018a] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. Understanding and mitigating the security risks of voice-controlled third-party skills on Amazon Alexa and Google Home. *arXiv preprint arXiv:1805.01525*, pages 1–16, 2018.

[Zhang *et al.* 2018b] Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Zhang *et al.* 2018c] Richong Zhang, Zhiyuan Hu, Hongyu Guo, and Yongyi Mao. Syntax encoding with application in authorship attribution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2742–2753, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[Zhang *et al.* 2019a] Hong Zhang, Hanshuo Wei, Yeye Tang, and Qiumei Pu. Research on classification of scientific and technological documents based on naive Bayes. In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, pages 327–331. ACM, 2019.

[Zhang *et al.* 2019b] Yang Zhang, Tianyuan Liu, Liqun Chen, Jinxurong Yang, Jiayi Yin, Yuncong Zhang, Zhixi Yun, Hao Xu, Lin Ning, Fengbiao Guo, et al. Riscoper: a tool for RNA–RNA interaction extraction from the literature. *Bioinformatics*, 2019.

[Zhao and Zobel 2005] Ying Zhao and Justin Zobel. Effective and scalable authorship attribution using function words. In *Asia Information Retrieval Symposium*, pages 174–189. Springer, 2005.

[Zhao and Zobel 2007] Ying Zhao and Justin Zobel. Searching with style: Authorship attribution in classic literature. In *ACM International Conference Proceeding Series*, volume 244, pages 59–68. Citeseer, 2007.

[Zhao *et al.* 2009] Yao Zhao, Yinglian Xie, Fang Yu, Qifa Ke, Yuan Yu, Yan Chen, and Eliot Gillum. BotGraph: Large scale spamming botnet detection. In *NSDI*, volume 9, pages 321–334, 2009.

[Zhao *et al.* 2014] Shilin Zhao, Yan Guo, Quanhu Sheng, and Yu Shyr. Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC bioinformatics*, 15(10):1–2, 2014.

[Zhao *et al.* 2015] Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen Zou. A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics*, volume 16, pages 1–10. Springer, 2015.

[Zhao *et al.* 2019] Zhenjie Zhao, Andrew Cattle, Evangelos Papalexakis, and Xiaojuan Ma. Embedding lexical features via tensor decomposition for small sample humor recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[Zheng and Jin 2022] Wanwan Zheng and Mingzhe Jin. A review on authorship attribution in text mining. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1584, 2022.

[Zheng and Jin 2023] Wanwan Zheng and Mingzhe Jin. A review on authorship attribution in text mining. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(2):e1584, 2023.

[Zheng *et al.* 2004] Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari. Feature selection for text categorization on imbalanced data. *SIGKDD Explor. Newsl.*, 6(1):80â€“89, jun 2004.

[Zheng *et al.* 2006] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the Association for Information Science and Technology*, 57(3):378–393, 2006.

[Zheng *et al.* 2013] Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. Deep learning for Chinese word segmentation and PoS tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 647–657, 2013.

[Zheng *et al.* 2017] Weiguo Zheng, Hong Cheng, Lei Zou, Jeffrey Xu Yu, and Kangfei Zhao. Natural language question/answering: Let users talk with the knowledge graph. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 217–226. ACM, 2017.

[Zhou and Zafarani 2020] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[Zhou *et al.* 2008] Bin Zhou, Jian Pei, and WoShun Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM Sigkdd Explorations Newsletter*, 10(2):12–22, 2008.

[Zhou *et al.* 2015] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis Lau. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.

[Zhou *et al.* 2018] Deyu Zhou, Yang Yang, and Yulan He. Relevant emotion ranking from text constrained with emotion relationships. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 561–571, 2018.

[Zhou *et al.* 2021] Qifeng Zhou, Xiang Liu, and Qing Wang. Interpretable duplicate question detection models based on attention mechanism. *Information Sciences*, 543:259–272, 2021.

[Zhu and P Xing 2010] Jun Zhu and Eric P Xing. Conditional topic random fields. 2010.

[Zhu *et al.* 2009] Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th annual international conference on machine learning*, pages 1257–1264, 2009.

[Zimmer 2013] B Zimmer. The science that uncovered jk rowlingâ€™s literary hocus-pocus. *The Wall Street Journal*, 2013.

[Ziwei and Chua 2019] Bernice Yeow Ziwei and Hui Na Chua. An application for classifying depression in tweets. In *Proceedings of the 2nd International Conference on Computing and Big Data*, pages 37–41, 2019.

[Zlatkova *et al.* 2018] Dimitrina Zlatkova, Daniel Kopev, Kristiyan Mitov, Atanas Atanasov, Momchil Hardalov, Ivan Koychev, and Preslav Nakov. An ensemble-rich multi-aspect approach for robust style change detection. *CLEF 2018 Working Nots of CLEF*, 2018.

[Zoph *et al.* 2016] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016.

[Zuo *et al.* 2019] Chaoyuan Zuo, Yu Zhao, and Ritwik Banerjee. Style change detection with feed-forward neural networks. In *CLEF*, 2019.

COLOPHON

This work is based on a research supported by the Council for Scientific and Industrial Research of South Africa (CSIR). Any opinion, findings and conclusions or recommendations expressed in this material are those of the author and therefore, the CSIR does not accept liability in regard thereto.

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyXat: https://bitbucket.org/amiede/classicthesis/. The author of this thesis has made many modifications to the `classicthesis` template to arrive at this final version

---

Regularized Deep Neural Network for Post-Authorship Attribution

Abiodun Modupe | ***February 12, 2024***
abiodunmodupe@gmail.com

Copyright © University of the Witwatersrand, Johannesburg, South Africa.