

https://dspace.org/

Optimizing credit risk model classification using bert deep learning model and vader techniques

Raliphada, Pfarelo

2026
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/50262

Raliphada, Pfarelo. (2026). Optimizing credit risk model classification using bert deep learning model and vader techniques [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/50262

Downloaded from a DSpace repository. https://dspace.org/

Optimizing credit risk model classification using bert deep learning model and vader techniques

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Pfarelo Raliphada
1593057

Supervised by Dr Micheal Olusanya
Co-supervisor Dr Seun Olukanmi



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG



Ethics Clearance Number: WCSAM-2024-31

A Research Project submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science, Data Science

Abstract

Credit risk assessment remains a fundamental component of financial risk management, traditionally relying on structured quantitative indicators such as credit scores, financial statements, and repayment histories. However, these approaches often overlook contextual and sentiment-driven signals embedded in unstructured textual data. This study investigates the optimization of credit risk classification through the integration of Natural Language Processing (NLP) techniques, specifically transformer-based models including BERT, RoBERTa, and LLaMA 3.2, alongside the rule-based sentiment analysis tool VADER. Financial news articles were collected using the NewsCatcher API over a one-year period and used as the primary source of unstructured textual data. Sentiment extracted from financial news is treated as an indirect indicator of credit risk–related signals rather than a direct borrower-level default prediction mechanism. Model performance was evaluated using Accuracy, Precision, Recall, and F1-score. Comparative analysis against traditional machine learning models revealed that XGBoost achieved the highest overall accuracy of 94%, followed by Logistic Regression at 92%. Among transformer-based models, BERT achieved 91% accuracy and RoBERTa 90%, while LLaMA 3.2 recorded 78%. Class-level performance further highlighted differences in contextual understanding. XGBoost achieved an F1-score of 0.97 for Positive sentiment and 0.88 for Neutral sentiment, demonstrating strong balance across classes. BERT achieved an F1-score of 0.79 for Negative sentiment but struggled with Neutral sentiment (0.62), whereas RoBERTa achieved a more balanced Neutral F1-score of 0.67. Logistic Regression performed competitively with a Neutral F1-score of 0.87. LLaMA 3.2 exhibited significant bias toward Positive sentiment and failed to effectively classify Negative and Neutral categories. The findings demonstrate that while classical models achieved superior overall accuracy, transformer-based models provide valuable contextual and semantic insights for sentiment-driven credit risk analysis. This research contributes to the advancement of NLP-enhanced financial risk modeling and establishes a foundation for future work addressing class imbalance, model interpretability, and hybrid modeling approaches.

Declaration

I, Pfarelo Raliphada, hereby declare the contents of this research to be my own work. It is submitted in partial fulfillment of the requirements for the degree of Master of Science in Data Science at the University of the Witwatersrand, Johannesburg. This work has not been submitted before for any degree or examination at any other university.

Dedication

This work is lovingly dedicated to my mother, Dorah Raliphada, who passed away in January 2025.

Your love, strength, and sacrifice shaped the person I am today. Though you are no longer here to witness this moment, your spirit continues to guide me. Every step I take is rooted in the foundation you built. This achievement is for you.

May your soul rest in peace.

Acknowledgements

First and foremost, I would like to convey my heartfelt appreciation to Dr. Micheal Olusanya, my supervisor, and Dr. Seun Olukanmi, my co-supervisor, for their invaluable guidance, support, and mentorship throughout this research journey. Their expertise, encouragement, and insightful feedback have played a crucial role in shaping this work and fostering my academic growth.

I also wish to express my genuine gratitude to the University of the Witwatersrand, especially the School of Computer Science and Applied Mathematics, for creating a motivating academic environment and providing the resources essential for conducting this study.

Particular thanks are due to the National Research Foundation (NRF) for the financial backing that made this research feasible. Their commitment to research and development is crucial in promoting innovation and academic excellence in South Africa.

To my family and friends, I appreciate your constant love, patience, and support. Your faith in me has been a pillar of strength during the most difficult moments of this academic journey.

This work belongs to all of you as much as it belongs to me. Thank you all.

Contents

Preface
Abstract i
Declaration ii
Dedication iii
Acknowledgements iv
Table of Contents v
List of Figures vii
List of Tables viii
Publications ix

1 Introduction **1**
1.1 Problem Statement 2
1.2 Research Aim and Objectives 3
1.3 Research Questions 3
1.4 Scope of the Study 4

2 Literature Review **5**
2.1 Theoretical Background 5
2.2 Related Works 7
2.3 Credit Risk in Deep Learning Models 12
2.4 Deep learning–based Natural Language Processing 13
2.4.1 Language Modeling 13
2.4.2 Encoder-Decoder Architectures 14
2.5 The Transformer Model 14
2.5.1 BERT 15
2.5.2 RoBERTa 16
2.5.3 LLAMA 17
2.6 Comparative Analysis of Models and Datasets 18
2.7 Literature Search and Screening Process 24
2.8 Data Analysis and Discussion 28
2.8.1 Data Analysis 28
2.8.2 Discussion 31
2.9 Conclusion 32

3 Methodology **34**

3.1 Data Collection and Preprocessing 35
3.2 Sentiment Classification 36

4 Data Analysis and Results **38**
4.1 Results 38
4.1.1 Performance Comparison 39
4.2 Ethical Considerations 44
4.3 Discussion 45

5 Conclusion and Future Work **47**
5.1 Conclusion of the Study 47
5.2 Limitations of the Study 48
5.3 Future Work 50

References **83**

A Appendix **84**
A.1 Model Evaluation Metrics 84
A.2 Dataset 84
A.3 Hyperparameter Settings 85
A.4 Supplementary Figures and Tables 85
A.5 Ethics Clearance 85
A.6 Published Conference Paper 85
A.7 Article submitted to Springer Journal: Artificial Intelligence Review (Under Review) 92

List of Figures

2.1 Figure depicting the BERT model architecture. The figure is adapted from Rustam et al. Rustam *et al.* [2024], who developed a classification method for lecturer expertise areas using BERT. 15
2.2 Figure depicting the RoBERTa model architecture. The figure is adapted from Kumari Kumari [2023], who described modifications made to BERT for NLP in the blog post RoBERTa: A Modified BERT Model for NLP. . . . 16
2.3 Figure depicting the LLaMA 3.2 model architecture. The figure is adapted from the blog post, From Vision to Edge: Meta's LLaMA 3.2 Explained Encord [2024], which provides an overview of its advancements and applications. 17
2.4 Bar chart comparing the number of included versus excluded articles after semantic filtering. 25
2.5 PRISMA flow diagram outlining the systematic review process, including identification, screening, eligibility assessment, and inclusion of studies. 27
2.6 Distribution and publication trends of selected studies over the years, illustrating the increasing academic interest in credit risk classification using NLP techniques. 27
2.7 Word cloud illustrating the most frequently occurring terms in the titles, abstracts, and keywords of the screened literature, highlighting key themes such as LLM, methods, models, classification, prediction, and risk. 28

3.1 A figure illustrating a random selection of 5 rows from the Training dataset 37

4.1 A figure illustrating a Sentiment Distribution Train vs Validation Data . . 38
4.2 Confusion matrix for BERT Model. 39
4.3 Confusion matrix for RoBERTa Model. 40
4.4 Confusion matrix for LLama 3.2 Model. 40
4.5 Confusion matrix for Logistic Regression Model. 41
4.6 Confusion matrix for XGBoost Model. 41

List of Tables

2.1 Selected Related Studies on NLP and Deep Learning Techniques Relevant to Text Analysis 8
2.2 Benchmark NLP Datasets Relevant to Text Classification and Sentiment Analysis 18
2.3 Model Overview from Reviewed Articles on NLP in Financial and Credit Risk Analysis 21

4.1 Model Performance Comparison 42
4.2 Performance Metrics for BERT, RoBERTa, LLaMA 3.2, Logistic Regression, and XGBoost 43
4.3 Confusion Matrix Summary for Each Model 43

Publications

Published Conference Paper

- **Published in:** IEEE EUROCON 2025 – 21st International Conference on Smart Technologies
- **Title:** *Optimizing Credit Risk Model Classification Using BERT, RoBERTa, and LLaMA 3.2*
- **Date of Conference:** 04–06 June 2025
- **Date Added to IEEE Xplore:** 15 July 2025
- **DOI:** 10.1109/EUROCON64445.2025.11073452
- **Publisher:** IEEE
- **Conference Location:** Gdynia, Poland

Article Submitted (Under Review)

- **Title:** *Transformer-based NLP Approaches for Credit Risk Prediction: A Systematic Review*
- **Type:** Research
- **Springer Journal:** Artificial Intelligence Review
- **Submission ID:** f8b0abc9-dc06-432c-8e2e-aee8e8376aac
- **Submission Version:** v.1.0
- **Current Status:** Submission has passed the technical checks

Chapter 1

Introduction

Financial institutions' top priorities are credit risk assessment and management, which makes complex model development and improvement necessary. The integration of cutting-edge technology, in particular Natural Language Processing (NLP), has shown promise in recent years as a means of improving credit risk model classification Lis *et al.* [2023] and Zhang *et al.* [2020]. NLP techniques have the ability to increase model accuracy, improve decision-making processes in the banking industry, and extract insightful information from textual data Başer *et al.* [2023]. Credit risk models are implemented using complicated methodology and data analysis approaches to anticipate default probabilities and assess borrower creditworthiness. Traditional approaches frequently rely on structured data, such as financial statements, credit scores, and payment history Ammann [2001]. However, the introduction of NLP allows for the incorporation of unstructured data sources such as customer correspondence, loan applications, and economic news into the modeling framework Rogojan *et al.* [2023]. A credit risk model refers to a computational or analytical framework used to assess the level of credit-related risk, while credit risk model classification involves categorizing such risk into predefined classes based on available structured and unstructured information.

NLP has evolved to include various techniques that can significantly enhance credit risk modeling. For instance, VADER (Valence Aware Dictionary and sentiment Reasoner) is a rule-based sentiment analysis model designed to capture sentiment in social media and other short text formats. It offers financial institutions a method to quantify sentiment from customer interactions and market news, which can be critical in assessing credit risk Hutto and Gilbert [2014]. Moreover, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a powerful tool in NLP, allowing for a deeper understanding of context and semantics in textual data. By employing BERT, financial institutions can enhance the accuracy of their credit risk models through improved sentiment classification and contextual analysis of unstructured data Devlin *et al.* [2018].

Additionally, LLaMA (Large Language Model Meta AI) models, particularly when fine-tuned for specific tasks such as aspect-based sentiment analysis (ABSA), showcase the potential for extracting nuanced sentiment information from text. Studies have indicated that fine-tuned LLaMA models can outperform previous state-of-the-art models

in ABSA tasks, emphasizing their effectiveness in identifying aspect terms, categories, and sentiment polarity in product reviews and social media posts Šmíd *et al.* [2024].

Financial institutions can use NLP to enhance their set of features, uncover sentiment patterns, and detect sophisticated risk indicators that standard numerical variables alone cannot capture Pandey *et al.* [2017]. This interdisciplinary approach is consistent with the changing environment of data-driven decision-making, in which the combination of quantitative models with qualitative insights strengthens the robustness and predictive power of credit risk assessments Li and He [2019]. Despite the potential benefits, the integration of NLP into credit risk modeling poses unique challenges and complexities. Issues such as data preprocessing, feature extraction, sentiment analysis accuracy, and model interpretability require careful consideration and methodological refinement Zamore *et al.* [2018]. Moreover, regulatory compliance, model transparency, and ethical considerations regarding data privacy and bias mitigation remain critical considerations in this context Kim and Kim [2014]. While these challenges are acknowledged, the scope of this study is limited to investigating sentiment-driven approaches for optimizing credit risk model classification using financial news data.

This study seeks to investigate the optimization of credit risk model classification by the strategic integration of NLP approaches. By investigating the synergy between quantitative modeling and textual analysis, this study hopes to improve credit risk models' prediction capacities, decision-making processes, and contribute to the growth of risk management techniques in the financial industry. In this context, sentiment extracted from textual data is treated as an indirect indicator of credit risk–related signals rather than a direct borrower-level default prediction.

1.1 Problem Statement

The accurate assessment of credit risk remains one of the most persistent challenges in financial analytics and banking. Conventional credit risk models have traditionally relied on structured data such as financial statements, credit scores, and historical repayment behavior to predict the probability of default Ammann [2001]. While these methods provide a strong quantitative foundation, they often fail to capture the contextual and sentiment-driven indicators embedded in unstructured data sources such as financial news, social media activity, and customer communications Rogojan *et al.* [2023], Lis *et al.* [2023].

Recent advances in Natural Language Processing (NLP), particularly transformer-based architectures like BERT, RoBERTa, and LLaMA, have shown substantial promise in enhancing credit-risk classification by incorporating text semantics, contextual embeddings, and sentiment polarity into financial prediction models Devlin *et al.* [2018], Liu *et al.* [2019b], Touvron *et al.* [2023]. Furthermore, VADER (Valence Aware Dictionary and Sentiment Reasoner) provides a complementary rule-based framework that allows for the quantification of sentiment in short-form text such as tweets or news headlines critical for capturing real-time market and consumer sentiment Hutto and Gilbert [2014].

Despite these advancements, several limitations continue to hinder the full integration of NLP in credit risk modeling. Key challenges include the complexity of data preprocessing, feature extraction from unstructured text, model interpretability, and the potential propagation of algorithmic bias Zamore *et al.* [2018]. Additionally, the ethical and regulatory dimensions of automated decision-making in finance such as transparency, fairness, and data privacy demand careful consideration to ensure responsible use of AI in credit assessments Kim and Kim [2014]. While these challenges are acknowledged, not all of them are empirically addressed within the scope of this study.

Given these challenges, this study seeks to explore how NLP-based sentiment and semantic analysis techniques can be systematically leveraged to optimize credit risk classification. By combining quantitative modeling with textual analysis, the research aims to enhance the predictive capacity, interpretability, and ethical soundness of credit risk models, thereby contributing to more robust and equitable decision-making frameworks in financial institutions Li and He [2019] and Başer *et al.* [2023]. The focus of this research is on evaluating sentiment-driven classification approaches using financial news data as an indirect indicator of credit risk–related signals, rather than on direct borrower-level credit risk prediction or regulatory compliance assessment.

1.2 Research Aim and Objectives

The aim of this study is to optimize credit risk model classification by integrating sentiment analysis and transformer-based Natural Language Processing models, specifically BERT, RoBERTa, and LLaMA 3.2, using financial news articles as a source of unstructured textual data.

Objectives of the study are to:

- To extract sentiment information from financial news articles using the VADER sentiment analysis technique.
- To evaluate the performance of transformer-based deep learning models in classifying sentiment-related credit risk signals from financial text.
- To compare the effectiveness of transformer-based models with traditional machine learning approaches for sentiment-driven credit risk classification.

1.3 Research Questions

- How does integrating social media sentiment and news articles enhance credit risk models?
- What methods improve the transparency of credit risk assessments?
- How can we ensure fairness and transparency in credit risk models?

1.4 Scope of the Study

This study focuses on the optimization of credit risk classification through the application of Natural Language Processing (NLP) and sentiment-analysis techniques. It specifically explores the integration of transformer-based NLP models, including BERT, RoBERTa, and LLaMA, together with the rule-based VADER sentiment analyzer, to improve the accuracy and interpretability of credit risk models Devlin *et al.* [2018], Liu *et al.* [2019b], Touvron *et al.* [2023], Hutto and Gilbert [2014].

The research relies primarily on secondary textual data such as financial news articles, customer feedback, and social-media content which offer rich contextual and sentiment information that traditional numerical variables often overlook Rogojan *et al.* [2023], Li and He [2019]. Within this framework, the study performs a comparative evaluation of the reviewed NLP models using established performance metrics, including accuracy, precision, recall, and F1-score, to assess their predictive strength in credit-risk classification tasks Başer *et al.* [2023].

Methodologically, the research adopts a Systematic Literature Review (SLR) approach to identify, synthesize, and critically analyze studies published between 2015 and 2025. This structured process ensures comprehensive coverage of NLP-based applications in credit-risk modeling and aligns with the PRISMA screening framework used in this research. Geographically, the review incorporates a global range of studies but places emphasis on applications relevant to the financial and banking sectors in both emerging and developed economies Heng and Subramanian [2022], Pathak *et al.* [2023].

It should be noted that the study does not involve primary data collection or the development of proprietary NLP architectures. Instead, it concentrates on evaluating and benchmarking existing approaches to identify best practices, methodological gaps, and future research opportunities in integrating NLP techniques into credit-risk assessment systems.

Chapter 2

Literature Review

2.1 Theoretical Background

The primary focus of financial institutions is to assess and manage credit risk, necessitating the development and enhancement of complex credit risk models. This literature review provides a comprehensive overview of how Natural Language Processing and Large Language Models are used for credit risk classification, highlighting key studies and developments in the field. Accurate credit risk classification is critical to ensure the stability and profitability of financial institutions within the broader scope of financial risk management. Traditionally, credit scoring systems have relied largely on structured data and statistical methods to evaluate a borrower's credibility. However, due to the rapid increase in unstructured financial text data, such as loan descriptions, transaction records, and customer communications, there is a growing recognition of the importance of utilizing these data to improve credit risk evaluations Wang *et al.* [2019].

Natural Language Processing (NLP) models, Especially those leveraging deep learning frameworks like Long Short-Term Memory (LSTM) networks and transformers, have enabled financial institutions to derive actionable insights from unstructured textual information. These models have shown significant improvements in classification tasks by capturing syntactic and semantic nuances in textual data, thus increasing the predictive accuracy of traditional credit risk models Jin and Zhang [2024]. For instance, attention-based LSTM networks have been used to analyze borrower behavior patterns in peer-to-peer (P2P) lending platforms, resulting in enhanced credit scoring performance Jin and Zhang [2024].

Recent studies have extended the application of transformer-based architectures, originally developed for natural language processing, to healthcare specifically for Alzheimer's disease risk assessment from speech transcripts demonstrating their effectiveness in early disease detection Roshanzamir *et al.* [2021]. By leveraging models such as BERT (Bidirectional Encoder Representations from Transformers), researchers have demonstrated the efficacy of NLP in extracting meaningful risk indicators from loan narratives Sanz-Guerrero and Arroyo [2025a]. The integration of temporal models, including

Temporal Convolutional Networks (TCN) and DilateFormer, with NLP techniques further refines classification accuracy by capturing temporal dependencies alongside textual semantics Shen and Wu [2025a].

Despite these enhancements, several challenges remain. The interpretability of deep learning models, concerns regarding privacy, and the integration of insights derived from NLP into regulatory compliant frameworks continue to impede the widespread adoption of these approaches Liu *et al.* [2021]. Furthermore, the current literature lacks a cohesive methodology for evaluating and benchmarking NLP-based credit risk classification models, leading to a fragmented understanding of their comparative effectiveness.

These enhancements and ongoing challenges highlight the evolving role of Natural Language Processing (NLP) in credit-risk modeling within financial analytics. Research consistently emphasizes how large language models and transformer architectures such as BERT, RoBERTa, and LLaMA can enhance traditional credit-scoring systems by improving contextual understanding and classification accuracy Devlin *et al.* [2018], Liu *et al.* [2019b], Touvron *et al.* [2023]. However, to fully utilize these benefits, it is necessary to address persistent issues related to model transparency, data governance, and evaluation standardization. Prior studies have noted that while transformer-based NLP systems deliver high predictive performance, their interpretability remains limited, creating barriers to responsible financial deployment Zamore *et al.* [2018], Kim and Kim [2014]. Overcoming these challenges is therefore critical to the effective and ethical use of NLP in credit-risk classification within real-world financial contexts Heng and Subramanian [2022], Li and He [2019].

The primary aim of this literature review is to examine how the integration of NLP techniques with credit-risk model classification has evolved and to determine how these developments contribute to improved risk management, credit-scoring accuracy, and decision-making in financial institutions. Specifically, this review investigates transformer-based NLP models such as BERT, RoBERTa, and LLaMA, together with the sentiment-based VADER framework, to understand how these approaches can enhance predictive accuracy and interpretability in financial risk modeling Hutto and Gilbert [2014], Başer *et al.* [2023]. The study therefore explores how NLP models have been applied to improve credit-risk classification, evaluates the benefits and challenges associated with incorporating unstructured textual data including financial news and social-media sentiment and assesses which NLP or hybrid models demonstrate the greatest promise for advancing credit-risk prediction Rogojan *et al.* [2023], Sanz-Guerrero and Arroyo [2025a]. In doing so, it seeks to identify the methodological, ethical, and interpretability challenges that remain unresolved and to provide insights into how they may be addressed through transparent and explainable AI frameworks Heng and Subramanian [2022], Li and He [2019].

This chapter therefore establishes the conceptual foundation for understanding the transformative potential of NLP and LLMs in credit-risk modeling while recognizing the institutional and regulatory constraints that must be overcome to ensure transparency, fairness, and ethical alignment in financial decision-making Zamore *et al.* [2018], Kim and Kim [2014].

2.2 Related Works

This section provides background information on modern deep learning-based NLP models, specifically BERT, RoBERTa, and LLaMA 3.2, within the specialized domain of credit risk classification. To begin with, we will outline essential NLP principles and elucidate how these sophisticated models effectively process text data, particularly in the context of financial analysis.

Credit risk classification is a critical task that involves evaluating the likelihood of default by individuals or entities based on their historical performance and contextual information. The complexity of this undertaking lies in the need to analyze multifaceted inputs, including both unstructured text data (such as credit histories, financial reports, and market sentiment) and structured numerical data (like credit scores and income levels) Lessmann *et al.* [2015]. In this landscape, models like BERT, RoBERTa, and LLaMA 3.2 stand out for their ability to extract actionable insights from vast amounts of unstructured text, enabling more informed decision-making Devlin *et al.* [2018]; Liu *et al.* [2019b]; Touvron *et al.* [2023].

BERT, an acronym for Bidirectional Encoder Representations from Transformers, leverages bidirectional context through the attention mechanism inherent to the Transformer architecture. Its pretraining objectives, which encompass Masked Language Modeling and Next Sentence Prediction, are crucial for capturing nuanced linguistic patterns essential for accurate credit risk assessment Devlin *et al.* [2018]. By analyzing relationships within text, BERT can identify subtle shifts in language that may indicate potential credit risks.

Building on BERT's successes, RoBERTa (Robustly Optimized BERT Pretraining Approach) optimizes the training methodology by omitting certain tasks, such as Next Sentence Prediction, and incorporating dynamic masking strategies. These enhancements boost predictive accuracy and strengthen the model's ability to handle complex and large-scale datasets, thereby providing a more robust solution for credit risk classification Liu *et al.* [2019b].

LLaMA 3.2 (Large Language Model Meta AI) is designed as a specialized model for deployment in low-resource environments. With a focus on efficiency, it employs optimized data selection and advanced tokenization techniques that enhance scalability for extensive NLP tasks while maintaining high accuracy levels. This capability is particularly important in credit risk analysis, where efficiently processing and interpreting large volumes of financial data is crucial Touvron *et al.* [2023].

Central to these transformer-based models is the attention mechanism, which enables them to assess the importance of various words within a given context effectively. While this feature significantly enhances performance, it also presents challenges in terms of memory resource demands, as these models can be computationally intensive. In conclusion, we will examine potential strategies and solutions aimed at alleviating these memory constraints, facilitating the effective use of these state-of-the-art NLP models in credit risk classification tasks.

Table 2.1: Selected Related Studies on NLP and Deep Learning Techniques Relevant to Text Analysis

Articles	Relationship between articles	Contributions	Gaps
Xu *et al.* [2025c]	introduces a novel end-to-end paradigm for clinical information extraction (CIE) from Chinese radiology reports, termed normalized clinical information extraction (NCIE).	Introduces normalized clinical information extraction (NCIE) for structured information from radiology reports.	The dataset primarily uses chest X-ray and CT reports, limiting broader anatomical coverage.
Lin and Liao [2024]	This article introduces a lexicon-based prompt method for financial sentiment analysis, addressing the limitations of existing domain-adaptation techniques.	It proposes a lexicon-based prompt method, incorporating dictionary information into a language model to improve accuracy.	The model is capable of guaranteeing the precision of both positive and negative information because the dictionary contains only that information for certain words.
Du *et al.* [2024]	This study presents a text-based measure of corporate ESG (Environmental, Social, Governance) risk exposure, utilizing a Word2vec-derived lexicon from 75 ESG reports and 100 news articles.	This paper introduces a text analysis method for measuring corporate ESG risk, using quarterly earnings call discussions.	Existing research focuses on the ultimate ESG performance, but lacks research on ESG risks and its measurement
Kirtac and Germano [2024]	The document explores the critical role of financial sentiment in market forecasting and investment strategies, emphasizing the advancements brought by LLMs in sentiment estimation.	The paper explains the distinction between BERT-type models like RoBERTa and FinBERT and GPT-type models such as GPT-4, OPT, and LLaMA.	There is a requirement for sentiment analysis that combines both speech and visual information.
Gao *et al.* [2022]	Presents a novel semantically weighted multi-criteria linear optimization classifier (SW-MCLOC) for predicting Chinese word formation patterns.	Introduces a semantically weighted multi-criteria linear optimization classifier (SW-MCLOC) for predicting Chinese word formation patterns.	Class-imbalance and potential noise in the dataset could affect the robustness of the model.

Articles	Relationship between articles	Contributions	Gaps
Corradi *et al.* [2022]	Demonstrates the application of NLP to extract mechanistic information from scientific literature for toxicology.	It offers a proof-of-concept that NLP supports information extraction for modern toxicology.	The NLP pipeline requires automation to unify biological events and link them to semantic ontology terms.
Hu *et al.* [2024]	This study presents a BERT-IDCNN-CRF model for entity extraction in R&D project abstracts.	Introduces a BERT-IDCNN-CRF model for scientific entity recognition in R&D project abstracts	Limitations in the annotated datasets due to the complexity of scientific texts.
Babaalla *et al.* [2024]	Presents a comparative analysis of various approaches for extracting UML class diagrams from natural language specifications.	Showcases the accuracy of recent systems involving NLP and machine learning techniques.	Existing methods for translating requirements to UML diagrams have shortcomings.
Li *et al.* [2025d]	Presents LLM4Jobs, an unsupervised framework utilizing LLMs for automated occupation extraction and standardization from job postings and resumes.	LM4Jobs introduces a framework for occupation extraction and standardization tasks, combining LLM-based summarization with embedding matching.	The actual dataset is limited in size, and the synthetic datasets may not consistently capture the subtleties of the real world.
Chagnon *et al.* [2024]	This paper introduces BERTeley, a Python package built on BERTopic, specifically designed for topic modeling of scientific articles.	BERTeley enhances topic modeling on scientific articles by addressing challenges posed by structural words and shared research goals.	Dilution of potential learnings due to stringent formatting and common words.
Shafee *et al.* [2025]	This study evaluates the performance of various Large Language Model (LLM)-based chatbots, including ChatGPT and GPT4all, in binary classification and Named Entity Recognition (NER) tasks for Open-Source Intelligence (OSINT) in cybersecurity.	Provides a comparative analysis of LLM-based chatbot performance in specialized CTI tasks, including binary text classification and Named Entity Recognition (NER)	There is a significant lack of comparative research specifically focused on open-source LLM-driven chatbots in the realm of OSINT-related CTI.

Articles	Relationship between articles	Contributions	Gaps
Shen *et al.* [2025]	Evaluates the applications of NLP in emergency department (ED) triage.	NLP has been applied to triage data to predict patient outcomes, such as the need for admission and critical care.	There is a necessity for future and external validation in various countries and triage systems.
Song *et al.* [2024b]	Presents a financial risk monitoring and early warning model integrating Long Short-Term Memory (LSTM) and Transformer deep learning architectures to address the complexities of modern financial markets.	It emphasizes the Transformer model's sensitivity to nonlinear relationships, increasing adaptability to financial market data.	Long-term dependencies and nonlinear characteristics in time-series data arise from the constantly changing nature of financial markets.
Lee and Lee [2024]	Presents a novel approach to construction risk assessments by developing a knowledge base utilizing BERT and graph models.	Uses NLP and graph models to extract knowledge from unstructured data and create a risk-assessment knowledge base.	The BERT model relies heavily on construction safety and accident casebooks, potentially limiting its broader applicability.
Cerchiello *et al.* [2022]	Investigates the integration of financial news and numerical data to enhance bank distress classification using a deep learning framework.	Processes textual data using 'Doc2Vec' and combines it with financial numerical data to create an input vector.	The model has limitations in how it processes news and converts them into vectors for classifying distress.
Zhang *et al.* [2021]	Presents a novel GeoSemantic2vec algorithm leveraging deep learning and natural language processing (NLP) to analyze urban flooding events	Computes geographic semantic properties of land in urban flooding areas.	Due to limited computing power, the spatial resolution is low, limiting semantic information.

Articles	Relationship between articles	Contributions	Gaps
Chuang *et al.* [2022]	Presents a two-stage method for structuring unstructured CT liver-tumor reports using Natural Language Processing (NLP) and interpretable machine learning.	Contributes a readable method, extracting feature keywords and discovering impact keywords using SHAP	Ensemble learning, meta learning, federated learning, solutions for data imbalance, and stratified K-folds cross-validation were not used in this research.
Jing *et al.* [2022]	Presents a text-mining method utilizing word embedding and deep learning to analyze chemical accident cases, focusing on risk assessment and management.	Extracts common trends and accident precursors through correlation analysis and text classification	The text classification model requires improvement for industrialization.
Maldonado *et al.* [2019]	Proposes an enhancement to SMOTE for high-dimensional data, which is relevant to improving classification performance in imbalanced datasets.	Introduces a novel oversampling method tailored for high dimensional imbalanced datasets by integrating K-means clustering and nearest neighbor.	Traditional oversampling techniques generate synthetic samples without considering data distribution in high-dimensional space, leading to overfitting.
ul Abideen *et al.* [2024]	Presents a novel framework for stock market price prediction and risk analysis, integrating Long Short-Term Memory (LSTM) networks and XGBoost.	The model merges stock values and Reddit data for stock price prediction.	Current systems often lack social media analysis and comprehensive risk assessment for investors.
López Espejel *et al.* [2023]	This study evaluates the reasoning capabilities of large language models (LLMs), specifically GPT-3.5, GPT-4, and BARD	Conducts a comprehensive evaluation of GPT-3.5, GPT-4, and BARD on seven reasoning tasks using eleven datasets.	Suggests that current LLMs struggle with tasks that demand deeper logical inference and understanding.
Arslan *et al.* [2024c]	Explores the integration of RAG with Large LLMs to extract political event information from media content	Introduces the Political Event Extraction system that combines RAG and LLMs to extract political events from news articles.	The study notes limitations such as the use of Llama2 over other potential LLMs.

Articles	Relationship between articles	Contributions	Gaps
Xia *et al.* [2022]	This study develops a knowledge graph of mobile payment platforms using deep learning to analyze risks and inform policy.	This research builds a knowledge acquisition model for mobile payment risk using a knowledge graph and unsupervised methods.	The knowledge graph model can't extract required triples well due to noise data.
Yuan and Wei [2024]	This study investigates the influence of sentiment consistency between credit rating experts and bank managers on bank bankruptcy and systemic risk, utilizing 716 reports from U.S. commercial banks.	Investigates the impact of tone consistency between bank managers and credit rating experts on bank risk.	Existing research overlooks the combined perspectives of bank managers and credit rating analysts regarding bank risks.
OU and ZHANG [2025]	This study introduces the Adversarial Unsupervised Domain Adaptation Network (AUDAN) for audit risk assessment, addressing the challenges posed by domain shifts in audit issue data.	Introduces an Adversarial Unsupervised Domain Adaptation Network (AUDAN) for audit risk assessment and explores domain shifts in audit issue data.	The study identifies a gap in addressing domain shifts in machine learning for audit risk assessment.
Xu and Li [2021]	Evaluates machine learning (ML) algorithms in predicting suicide risk.	Examine ML outcomes for predicting suicide and provide a list of suicide risk factors.	ML requires sufficient data to train and validate early identification of risk factors.

2.3 Credit Risk in Deep Learning Models

Credit risk models are utilized to evaluate the potential financial loss that a credit institution, such as a bank or peer-to-peer lender, may suffer if a borrower defaults on their loan repayment. The most critical aspect of a credit risk model is the probability of default, which is typically estimated through statistical methods using credit scoring models. Pathak *et al.* [2023] examines how banks can adapt to manage credit risk amid economic uncertainty caused by rising interest rates, inflation, and geopolitical tensions. Heng and Subramanian [2022] provides an in depth examination of how machine learning enhances credit risk assessment and the role of Explainable AI in addressing its limitations. It highlights the high performance of machine learning al-

gorithms in predicting credit risk, while emphasizing their lack of interpretability. The review discusses Explainable AI techniques like LIME and SHAP, which improve transparency by explaining model decisions, thereby fostering trust and ensuring fairness in lending practices.

Jeyakarthic and Ramesh [2023] introduces the GPDBN-CRA model as a tool to assist financial institutions in evaluating loan applications. This innovative model integrates a Dynamic Bayesian Network to classify credit risk based on customer data, along with Genetic Programming for hyperparameter tuning, which enhances overall performance. Genetic programming is then used to optimize the parameters, leading to improved prediction accuracy. Hassija *et al.* [2020] proposes a novel approach to credit scoring by integrating blockchain technology and prospect theory. Blockchain ensures a secure, transparent, and tamper-proof credit scoring system, reducing reliance on centralized credit bureaus and enhancing data security. This innovative approach offers a secure, efficient method for assessing credit risk, potentially improving lending outcomes and minimizing defaults.

Jukna [2022] examines different credit scoring models to help financial institutions select the best method. It compares traditional techniques like logistic regression with advanced machine learning models, evaluating their accuracy, interpretability, and efficiency. The study highlights the practical implications of these models for various credit risk scenarios and explores trends like big data and explainable AI to improve transparency and trust in credit scoring.

2.4 Deep learning–based Natural Language Processing

Natural language processing is centered on enabling computers to understand and process human languages. These models are crafted to learn a language's structure, syntax, and linguistic patterns or specific tasks by training on extensive text datasets relevant to the target language and objective Brown *et al.* [2020].

The training of such models typically consists of three main steps. First, the text is segmented into smaller, manageable units through a process known as tokenization. These tokens are then transformed into a numerical format referred to as word embeddings, allowing for a more nuanced representation that supports mathematical operations Devlin *et al.* [2018]. Finally, a language model is trained using this tokenized and embedded dataset Brown *et al.* [2020].

2.4.1 Language Modeling

A language model is a statistical representation that quantifies the probabilities of sequences of words or tokens. By analyzing a corpus of linguistic data, it assigns likelihoods to potential subsequent sequences $(t_1, t_2, \ldots, t_n)$ in a particular language L. In essence, it acts as a predictive model that identifies the most probable next words in a given context Brown *et al.* [2020].

$$P(t_1, t_2, \ldots, t_n) = p(t_1)p(t_2|t_1) \cdots p(t_n|t_1, t_2, \ldots, t_{n-1})$$

$$= \prod_{i=1}^{n} p(t_i|t_1, t_2, \ldots, t_{i-1}) \tag{2.2}$$

Training a language model involves utilizing a large unlabeled dataset to establish a foundational understanding of language. Subsequently, these models can be fine-tuned for specific downstream tasks using labeled data. The classical language modeling objective focuses on learning a left-to-right context, predicting the conditional probability of all preceding tokens $t_1, t_2, \ldots, t_{i-1}$ for a given token t_i. This traditional approach, which only generates predictions based on previously encountered words, is known as auto-regressive or unidirectional. It is particularly well-suited for tasks related to language generation Brown *et al.* [2020].

A prevalent learning objective in natural language processing is the masked language model (MLM) objective. In this method, certain words or tokens within a sentence are randomly masked, and the model's goal is to predict these omitted tokens. By examining the context surrounding the masked words, the model assigns probabilities to the most likely candidates for replacement. This language modeling technique enables the model to learn in both left-to-right and right-to-left directions, facilitating what is known as bidirectional reasoning. This capability is particularly advantageous for sentence-level tasks, including text classification, named entity recognition, sentence analysis, and question answering Devlin *et al.* [2018].

2.4.2 Encoder-Decoder Architectures

The Encoder-Decoder architecture is a neural network design for sequence modeling, particularly in NLP tasks like sentiment analysis and text classification, where input lengths can vary while outputs remain fixed. This architecture transforms input sequences of arbitrary length into output sequences Pakdaman *et al.* [2025]. Traditional models struggle with this variability, requiring fixed lengths for training. The Encoder-Decoder consists of two networks: an encoder that converts an input sequence into a fixed-length vector representation, and a decoder that generates the output sequence from this vector. Both networks are trained together to accurately encode and reconstruct the target sequence.

2.5 The Transformer Model

Several pre-trained Transformer-based models, including LLaMA 3.2 Touvron *et al.* [2023], RoBERTa Liu *et al.* [2019b], and BERT Devlin *et al.* [2018], have set new standards for performance in various NLP applications since their introduction. RoBERTa enhances BERT's architecture by employing improved training methods, while LLaMA 3.2 focuses on efficiency and scalability in low-resource settings. The success of these

models is largely attributed to their effective use of the attention mechanism, which enables strong and context-aware language understanding.

2.5.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a language representation model created by researchers at Google AI Language in 2018. BERT was developed using deep learning techniques in conjunction with various other approaches. The architecture of the BERT model features a multi-layer bidirectional Transformer, which mirrors the original implementation of the Transformer but exclusively employs the encoder component Rustam *et al.* [2024].

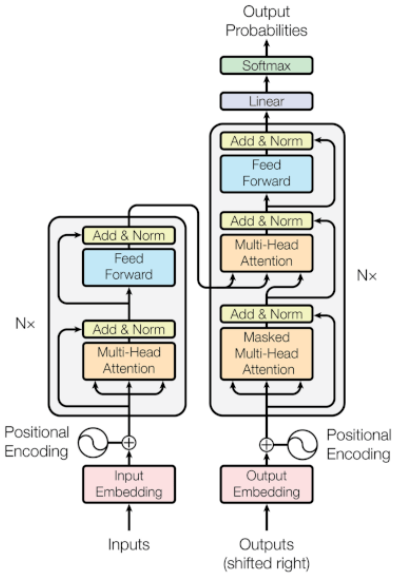

Figure 2.1: Figure depicting the BERT model architecture. The figure is adapted from Rustam et al. Rustam *et al.* [2024], who developed a classification method for lecturer expertise areas using BERT.

Koroteev [2021] highlight that BERT has had a transformative effect on NLP, showcasing its effectiveness in various tasks such as text classification, sentiment analysis, named entity recognition, and question answering. It consistently surpasses earlier models like GPT and ELMo. The review also addresses proprietary BERT-based models created for specific applications, highlighting BERT's significance in establishing new benchmarks in NLP and inspiring further advancements within the field. BERT is designed to pre-train deep bidirectional representations from unlabeled text by simultaneously taking into account both the left and right contexts across all layers. Each sequence consistently begins with a special classification token. The final hidden state corresponding to this token acts as a unified sequence representation for classification tasks. Pairs of sentences are merged into a single sequence Chotirat and Meesad [2021]. The study by Mann *et al.* [2023] explores the use of an Enhanced BERT model to enhance sentiment analysis of Twitter data. Utilizing the Kaggle SMILE dataset, the

research tailors BERT to accommodate the informal nature of tweets, which frequently incorporate slang, acronyms, and emoticons. The Enhanced BERT model achieved an impressive accuracy rate of 96% in discerning sentiments such as happiness and sadness, highlighting its effectiveness. This research emphasizes the potential of Enhanced BERT in precisely analyzing the sentiments expressed in short, and informal texts.

2.5.2 RoBERTa

RoBERTa is a an advanced NLP model developed by Facebook AI in 2019, building upon the BERT architecture with a series of enhancements designed to boost its performance Chen *et al.* [2023a].

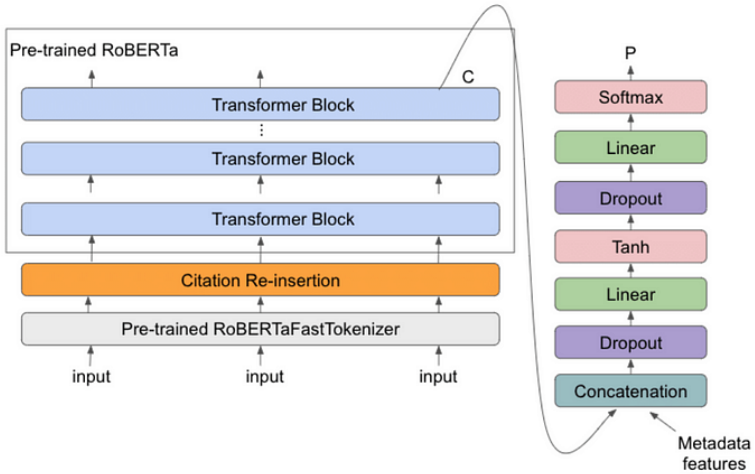

Figure 2.2: Figure depicting the RoBERTa model architecture. The figure is adapted from Kumari Kumari [2023], who described modifications made to BERT for NLP in the blog post RoBERTa: A Modified BERT Model for NLP.

Trained on an extensive range of comprehensive datasets, RoBERTa demonstrates outstanding results across various NLP tasks, including question-answering, sentiment analysis, and entity recognition. Its robustness and versatility have fostered its widespread application, making it a foundational model for fine-tuning and transfer learning in numerous NLP applications.

Liu *et al.* [2019b] explore how to enhance BERT's pretraining by modifying hyperparameters and increasing the training data size. Their findings indicated that BERT was not fully trained and showed that RoBERTa, utilizing better training methods, could outperform later models. RoBERTa set new records on benchmarks like GLUE, RACE, and SQuAD, highlighting the importance of hyperparameter adjustment and sufficient training data for improving language model performance. Alrowais *et al.* [2024] focuses on enhancing cyberbullying detection by combining RoBERTa's transformer capabilities with GloVe's word embeddings. This hybrid model improves understanding of abusive language by leveraging RoBERTa's contextual learning and GloVe's semantic representation. Training on social media datasets aims to effectively identify nuanced

instances of cyberbullying. Advanced models like RoBERTa are expected to achieve higher accuracy, supporting efforts to create safer online environments.

2.5.3 LLAMA

The LLaMA $3.2 - 1B$ model, part of the Large Language Model Meta AI (LLaMA) series, serves as a powerful tool for a variety of natural language processing tasks. With its 1 billion parameters, it strikes a balance between computational efficiency and performance, making it ideal for both research and practical applications, particularly in settings with limited resources. Optimized for tasks such as text generation, summarization, and question answering, LLaMA $3.2 - 1B$ delivers impressive results while maintaining a lightweight profile compared to larger models in the series. Its design is tailored for users seeking an effective model that does not demand extensive computational power Hugging Face [2024].

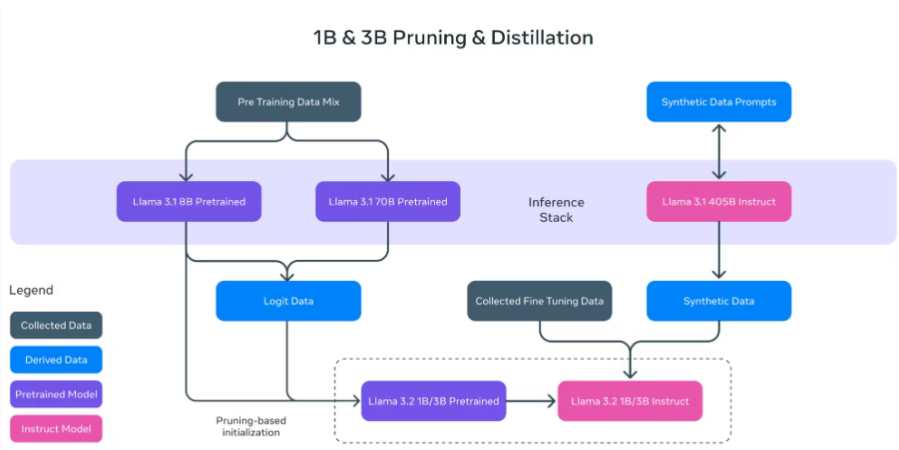

Figure 2.3: Figure depicting the LLaMA 3.2 model architecture. The figure is adapted from the blog post, From Vision to Edge: Meta's LLaMA 3.2 Explained Encord [2024], which provides an overview of its advancements and applications.

Stefan and Brad [2024] highlights the exceptional performance of LLaMA 3 in identifying subtle misinformation. It outperformed state-of-the-art competitors, achieving superior accuracy and precision because of its advanced contextual understanding and adaptability to complex narratives. Additionally, LLaMA 3 demonstrated remarkable efficiency in processing large datasets, rendering it suitable for real-time applications. Notably, the model demonstrated reasoning capabilities similar to human judgment, significantly reducing both false positives and false negatives. These findings emphasize its potential as a powerful tool for combating misinformation. Sarfati *et al.* [2024] discusses how large language models (LLMs) emulate human cognitive processes like reasoning and decision-making. It highlights advancements such as multi-layer attention mechanisms and transformer networks that enable effective handling of complex language tasks. The paper also addresses the importance of improving model interpretability and considers ethical issues related to bias, fairness, and misuse, while em-

phasizing LLMs' transformative potential in sectors like healthcare, finance, and education.

2.6 Comparative Analysis of Models and Datasets

Table 2.2 outlines the datasets employed in the reviewed studies. These datasets span from unstructured sources like financial tweets and news articles to structured stock market time series and legal judgments. Most are annotated with sentiment or emotion labels, while others use inferred labels like risk tolerance or psychometric traits. A few works rely on synthetic or proprietary data, limiting reproducibility and cross-study comparability. The table categorizes sources by type, language, size (where available), and annotation strategy, providing insight into the diversity and challenges of data collection in financial NLP. It also clarifies each dataset's intended analytical purpose, such as credit scoring, supplier evaluation, or emotion classification.

Table 2.2: Benchmark NLP Datasets Relevant to Text Classification and Sentiment Analysis

Source	Data Type	Language	Annotation	Purpose
P2P Lending Descriptions [Yuan and Wei 2024]	Unstructured loan text	English	No manual annotation; LLM-derived risk indicators	Construct interpretable risk indicators from borrower-written descriptions using Large Language Models
Social Media Posts [Yang *et al.* 2022]	User-generated text from P2P lending platforms	English	Self-supervised psychometric labeling	Infer psychological traits for credit risk assessment
P2P Loan Applications [Brahma *et al.* 2021]	User-generated loan descriptions	English	Labeled with default outcomes	Predict credit default using deep learning on textual data
Local Government Work Reports [Guo *et al.* 2022]	Government documents	Chinese	Sentiment scores derived via NLP	Assess regional debt risks through sentiment analysis of official reports
Bank Financial Data [Tao *et al.* 2025a]	Structured financial indicators	Not specified	Labeled with default outcomes	Predict financial risks using deep learning models
Policy Documents on Mobile Payments [Xia *et al.* 2022]	Government policy documents	Chinese	Unsupervised extraction of entities and relationships	Construct knowledge graphs for risk analysis and policy implications in mobile payment platforms

Source	Data Type	Language	Annotation	Purpose
Financial News Articles [Adhikari *et al.* 2023]	Financial news headlines	English	Labeled with sentiment polarity	Enhance sentiment analysis using hybrid word representations and explainability techniques
P2P Platform Data [Fu *et al.* 2020]	Platform information and investor comments	Chinese	Labeled with platform default outcomes	Evaluate default risk of P2P platforms from investors' perspectives using deep learning
Financial Text Data [Lin and Liao 2024]	Financial documents	Not specified	Annotated with sentiment labels using lexicon-based prompts	Enhance financial sentiment analysis through lexicon-based prompt methods
Legal Judgments [Wang *et al.* 2025c]	Legal case documents related to supply chain finance	Chinese	Annotated with credit risk outcomes	Utilize legal judgments to improve credit risk prediction in supply chain finance
Financial Twitter Data [García-Méndez *et al.* 2023]	Tweets mentioning financial assets	English	Annotated with 'opportunity' and 'precaution' emotions per asset	Detect asset-specific financial emotions for market screening and decision-making
User-Generated Texts [Xing 2024]	Text data from various sources	English	Labeled with financial risk tolerance levels	Profile financial risk tolerance using textual data
Scientific Article Corpora [Chagnon *et al.* 2024]	Scientific research articles	English	Unsupervised; evaluated using coherence metrics	Benchmark topic modeling performance on scientific literature using BERTeley framework
Financial Reports and Market Data [Che *et al.* 2024]	Annual reports, financial indicators, and market data	English	Labeled with financial distress outcomes	Predict financial distress using an attentive and interpretable multimodal framework
Financial Transaction Graphs [Boyapati and Aygun 2025]	Transactional data represented as graphs	English	Labeled with fraudulent and non-fraudulent classes	Detect fraud in imbalanced datasets using graph neural networks

Source	Data Type	Language	Annotation	Purpose
NLI Benchmark Datasets [Kim *et al.* 2024]	Sentence pairs for entailment tasks	English	Labeled with entailment, contradiction, or neutral	Enhance NLI performance using syntax extension with virtual linking-phrases
Job Postings and Resumes [Li *et al.* 2025d]	Free-text job descriptions and resumes	English	Unsupervised; evaluated against standard occupation taxonomies	Extract and standardize occupations using LLMs for improved job matching and labor market analysis
Twitter OSINT Dataset [Shafee *et al.* 2025]	Tweets related to cybersecurity threats	English	Labeled for binary classification and named entity recognition tasks	Evaluate LLM-based chatbots for cyber threat detection and information extraction
Policy Texts [Liang *et al.* 2022]	Government policy documents	English	Keywords labeled for extraction tasks	Address class imbalance in keyword extraction using ensemble oversampling with three-way decisions and SMOTE
Annual Reports [Wang *et al.* 2023]	Financial statements and textual disclosures	English	Labeled for fraud detection	Detect financial statement fraud using multimodal data with fine-grained attention mechanisms
Financial Disclosures [Yuan and Wei 2024]	Annual reports and financial statements	English	Labeled for sentiment consistency and bank risk levels	Analyze the impact of sentiment consistency on bank risk using deep learning models
Price and Financial News Data [Du and Tanaka-Ishii 2022]	Structured price data and unstructured financial texts	English	Embedded using NESTED (News-STock Event Distribution) framework	Integrate price and textual event data to optimize portfolios with improved tail risk sensitivity
ESG Reports and Financial Texts [Du *et al.* 2024]	Financial reports and ESG disclosures	English	Annotated for ESG risk factors using text analysis	Develop a text-based measure to assess firm-level ESG risk exposure and its impact on financial performance

Table 2.3 presents a comparative overview of NLP models applied in financial and credit risk analysis. The models range from transformer-based architectures to convolutional and hybrid deep learning structures, each tailored to a specific problem such as sentiment analysis, credit scoring, or entity disambiguation. Most architectures integrate

attention mechanisms or hybrid word embeddings to address polysemy, interpretability, and contextual representation in financial text. Some models, such as PsyCredit and PLWL, incorporate explainability techniques like LRP or dictionary filtering to enhance transparency. Performance across models varies, often based on task and dataset, but many demonstrate significant improvement over traditional baselines. The table also highlights which models emphasize explainability, accuracy, or application novelty.

Table 2.3: Model Overview from Reviewed Articles on NLP in Financial and Credit Risk Analysis

Model	**Architecture**	**Data Source**	**Input Features**	**Performance**
LLM Risk Indicator [Yuan and Wei 2024]	Zero-shot LLM (GPT-3.5/BERT baseline)	P2P loan descriptions	Borrower-written loan purposes and intentions	LLM-generated risk score correlated with actual default, outperforming lexical baseline
PsyCredit [Yang *et al.* 2022]	Deep learning (LRP)	Social media (P2P lending)	Psychometric signals from language	Outperformed baseline credit models
Deep Learning Credit Default Predictor [Brahma *et al.* 2021]	Deep neural network (DNN)	P2P loan application texts	Textual features from loan descriptions	Improved prediction accuracy over traditional models
LocalGov Debt Risk Model [Guo *et al.* 2022]	Deep learning with NLP and sentiment analysis	Local government work reports	Sentiment scores related to debt risks	Identified high-risk regions
grcForest_XGB [Tao *et al.* 2025a]	Ensemble model combining grcForest and XGBoost	Bank financial data	Structured financial indicators	Outperformed other models in risk prediction
MobilePay Risk KG [Xia *et al.* 2022]	Deep learning-based unsupervised knowledge graph construction	Policy documents on mobile payments	Extracted entities and relationships from text	Enabled risk analysis and policy implications through knowledge graph visualization
Hybrid Word Representation Model [Adhikari *et al.* 2023]	CNN with attention mechanism using hybrid embeddings	Financial news articles	Combined word embeddings capturing semantics, syntax, and polysemy	Outperformed baseline models

Model	Architecture	Data Source	Input Features	Performance
Investor-Centric P2P Risk Model [Fu *et al.* 2020]	ELMo-BiLSTM-CNN-CRF sequence labeling	P2P platform data	Extracted keywords and investor sentiments	Enhanced default risk prediction of P2P platforms
Lexicon-Based Prompt Model [Lin and Liao 2024]	Pre-trained language model with lexicon-based prompt tuning	Financial text data	Financial lexicon terms and prompt templates	Improved sentiment analysis accuracy
ACWGAN-GPSA [Wang *et al.* 2025c]	Attention-based Conditional Wasserstein GAN with Gradient Penalty and Self-Attention	Legal judgments in supply chain finance	Legal case features, clauses, and risk indicators	Demonstrated superior performance in credit risk prediction tasks
TABEA [García-Méndez *et al.* 2023]	Clause segmentation + emotion tagging	Financial Twitter data	Tweet clauses, asset mentions, emotion types	Over 90% accuracy
CNN Risk Profiler [Xing 2024]	CNN + psycholinguistic signals	Synthetic user-generated text	Linguistic markers + embeddings	Micro-F1 0.51
BERTeley [Chagnon *et al.* 2024]	Transformer-based topic modeling with pre-trained scientific language models	Scientific article corpora	Sentence embeddings from scientific texts	Demonstrated superior coherence and topic quality
Multimodal Financial Distress Predictor [Che *et al.* 2024]	Attentive multimodal deep learning framework	Financial reports and market data	Textual and numerical financial features	Achieved superior performance in predicting financial distress
BalancerGNN [Boyapati and Aygun 2025]	Graph Neural Network with balancing techniques	Financial transaction graphs	Node features and graph structures	Improved fraud detection performance on imbalanced datasets
Virtual Linking-Phrase (VLP) Framework [Kim *et al.* 2024]	Syntax-augmented transformer model	NLI benchmark datasets	Sentence pairs with syntactic virtual linking-phrases	Improved accuracy

Model	Architecture	Data Source	Input Features	Performance
LLM4Jobs [Li *et al.* 2025d]	Unsupervised LLM-based framework	Job postings and resumes	Raw textual data	Outperformed existing unsupervised methods
LLM-based Chatbots for OSINT [Shafee *et al.* 2025]	Various LLMs including ChatGPT, GPT4all, Dolly, Stanford Alpaca, Alpaca-LoRA, Falcon, and Vicuna	Twitter OSINT Dataset	Textual data from tweets	GPT-4 achieved an F1 score of 0.94 in binary classification; GPT4all achieved 0.90
Ensemble Oversampling with 3WD and SMOTE [Liang *et al.* 2022]	Bagging with logistic regression, three-way decisions, and SMOTE	Policy texts	Textual features for keyword extraction	Improved F-measure and G-mean
RCMA [Wang *et al.* 2023]	Attention-based multimodal deep learning model	Annual reports	Financial ratios and textual disclosures	Outperformed state-of-the-art benchmarks in financial statement fraud detection tasks
Sentiment Consistency Model [Yuan and Wei 2024]	Deep learning model analyzing sentiment consistency	Financial disclosures	Sentiment scores and consistency metrics	Enhanced risk assessment accuracy
NESTED Portfolio Optimizer [Du and Tanaka-Ishii 2022]	Hilbert space embedding of price and news text vectors	Stock prices and financial news articles	Price signals and text-based event embeddings	Achieved up to 45.5% higher returns and 59.4% Sharpe ratio improvement over baseline models
ESG Risk Text Analysis Model [Du *et al.* 2024]	NLP techniques applied to ESG-related textual data	Corporate ESG reports and financial statements	ESG-related terms and phrases extracted from texts	Demonstrated that the text-based ESG risk measure is a significant predictor of firms' financial risk and performance

2.7 Literature Search and Screening Process

This study uses a Systematic Literature Review (SLR) methodology to systematically identify, evaluate, and synthesize academic research on the integration of Natural Language Processing (NLP) models, specifically transformer-based architectures such as BERT, RoBERTa, and LLaMA, in credit risk classification tasks. The SLR uses Python based tools to ensure methodological transparency and reproducibility by integrating structured database querying with automated processes for filtering, cleaning, and scoring semantic similarity. To ensure comprehensive coverage, scholarly articles were sourced from three prominent databases: Scopus, ScienceDirect, and Web of Science. A standardized Boolean search string was crafted to explore these sources, focusing on papers on credit risk, financial risk, or loan default classification utilizing NLP or text mining techniques. The search was confined to English language publications released between 2015 and 2025. The query structure for Scopus is as follows:

```
TITLE-ABS-KEY (
  (credit risk OR financial risk OR loan default OR credit scoring)
  AND (natural language processing OR NLP OR text mining)
  AND (classification OR prediction)
)
AND PUBYEAR > 2014 AND PUBYEAR < 2026
AND (
  EXCLUDE(DOCTYPE, cp) OR EXCLUDE(DOCTYPE, re) OR
  EXCLUDE(DOCTYPE, cr) OR EXCLUDE(DOCTYPE, sh) OR
  EXCLUDE(DOCTYPE, ed)
)
AND (EXCLUDE(LANGUAGE, Chinese))
```

ScienceDirect's query was formulated using the basic search bar and fine-tuned for optimal keyword matching. Although the interface lacks the Boolean complexity found in platforms like Scopus, it allows for a direct logical structure that effectively yields relevant peer-reviewed results. The search targeted the comprehensive full-text repository of Elsevier publications to locate articles that discuss NLP in the context of credit risk. The logical structure emphasized classification and prediction, specifically incorporating credit-related terminology. The search was also limited to publications in the English language that were released between 2015 and 2025. The following query demonstrates the syntax compatible with ScienceDirect's search functionality:

```
(credit risk OR financial risk OR loan default OR credit scoring) AND
(natural language processing OR NLP OR text mining) AND
(classification OR prediction)
AND publicationYear:[2015 TO 2025] AND language:English
```

The Web of Science platform offers a Topic Search (TS) feature for thorough querying of article titles, abstracts, author keywords, and Keywords Plus. The final query structure has been refined for universal compatibility and is implemented through the core collection interface. To narrow down the results, only publications from 2015 to 2025 in English were considered, and document types such as reviews and editorial materials were excluded. Below is the corresponding query that adheres to Web of Science's default logic syntax:

```
TS=(credit risk OR financial risk OR loan default OR credit scoring)
   AND
TS=(natural language processing OR NLP OR text mining) AND
TS=(classification OR prediction)
AND PY=(2015-2025)
AND LA=(English)
NOT DT=(Meeting Abstract OR Editorial Material OR Review)
```

The initial search produced 284 unique records across the databases, which were then exported in BibTeX format and consolidated into a single bibliography. Following this export, a semantic filtering process was implemented to evaluate the relevance of each article against a predefined research theme. Of the total records, 217 articles were excluded due to low semantic similarity, while 63 articles met the inclusion criteria. A comparative bar chart was created to visualize the proportions of included versus excluded articles, providing a clear representation of the data refinement process that underlies the methodological rigor of the review.

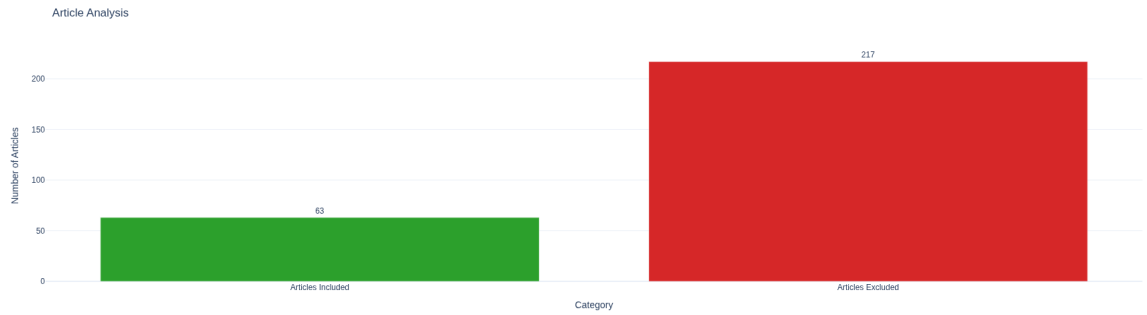

Figure 2.4: Bar chart comparing the number of included versus excluded articles after semantic filtering.

A Python script utilizing the `bibtexparserlibrary` was used to load and parse the entries, while duplicates were identified and eliminated by string normalization of titles and unique identifiers.

```
import bibtexparser
from bibtexparser.bparser import BibTexParser
from bibtexparser.customization import homogenize_latex_encoding

def normalize(text):
    return text.strip().lower().replace('\n', '') if text else ''

def remove_duplicates(entries):
    seen_titles = set()
    unique_entries = []
    for entry in entries:
        title = normalize(entry.get(title, ""))
        if title not in seen_titles:
            seen_titles.add(title)
            unique_entries.append(entry)
    return unique_entries
```

After deduplication, 280 unique records remained. To refine the corpus to topically relevant studies, a semantic filtering step was introduced. Using the `SentenceTransformer` model `all-MiniLM-L6-v2`, cosine similarity was calculated between the embedding of each article generated from its title, and keywords and a target embedding representing the review's central theme:

```
from sentence_transformers import SentenceTransformer, util
import torch

model = SentenceTransformer('all-MiniLM-L6-v2')

target_topic = (
 Optimising credit risk model classification using NLP,
    BERT, RoBERTa, and LLaMA
)
target_embedding = model.encode(
    target_topic, convert_to_tensor=True
)

SIMILARITY_THRESHOLD = 0.35
similar_articles = []

for entry in unique_entries:
    combined_text = "".join([
        entry.get(title, ""),
        entry.get(keywords, "")
    ])
    if not combined_text.strip():
        continue

    entry_embedding = model.encode(
        combined_text, convert_to_tensor=True
    )
    similarity = util.cos_sim(
        target_embedding, entry_embedding
    ).item()

    if similarity >= SIMILARITY_THRESHOLD:
        entry["similarity"] = round(similarity, 3)
        similar_articles.append(entry)
```

Only articles that scored above a semantic similarity threshold of 0.35 were selected for detailed analysis, resulting in a curated collection of 63 studies. This filtering process ensured that the reviewed literature was closely aligned with the thematic focus on credit risk modeling utilizing NLP. To uphold methodological transparency, the screening process adhered to PRISMA principles. The key phases included article identification, deduplication, semantic evaluation, and final inclusion based on relevance. Each step contributed to refining the dataset into a robust evidence base. The distribution and flow of articles across these stages are visually represented in the Sankey-style PRISMA diagram provided below.

A bar chart illustrating publication trends reveals a distinct upward trajectory in relevant research output, particularly from the year 2020 onward. This pattern underscores

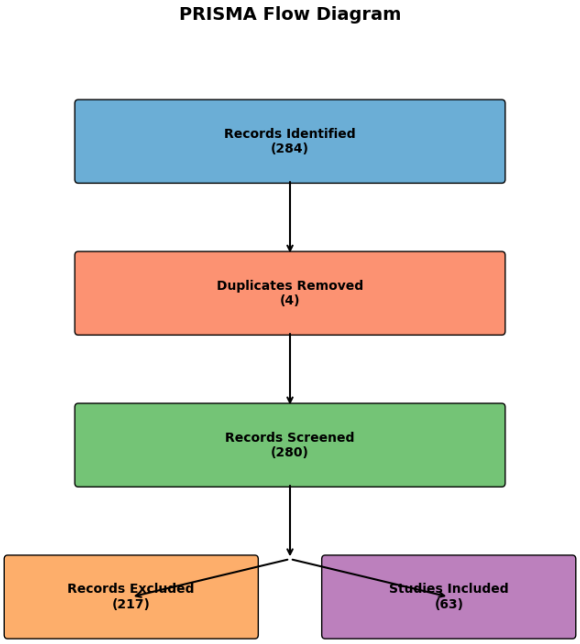

Figure 2.5: PRISMA flow diagram outlining the systematic review process, including identification, screening, eligibility assessment, and inclusion of studies.

a growing scholarly focus on the application of NLP in financial risk modeling. The peak observed in 2025 indicates a recent surge in academic engagement, likely influenced by advancements in transformer-based models and the increasing availability of financial text data. This trend also mirrors the broader adoption of machine learning in finance, coinciding with various regulatory and technological developments. Overall, the temporal analysis emphasizes the timeliness and significance of this study within the evolving research landscape.

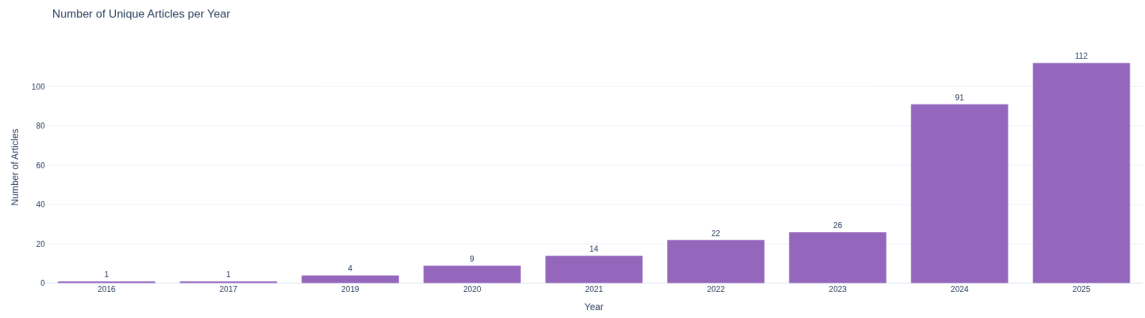

Figure 2.6: Distribution and publication trends of selected studies over the years, illustrating the increasing academic interest in credit risk classification using NLP techniques.

Additionally, a word cloud was created by aggregating text from article titles, abstracts, and keywords provided by the authors. This visualization technique allowed for the identification of key thematic patterns within the filtered corpus. Notable terms such as "LLM," "classification," "risk," and "prediction" emerged with high frequency, highlighting the central focus areas of the studies included. The prominence of these keywords emphasizes the significance of large language models and predictive analytics within the realm of credit risk modeling. This thematic consistency further reinforces the soundness of the article selection process and supports the alignment of the literature with the study's research objectives.

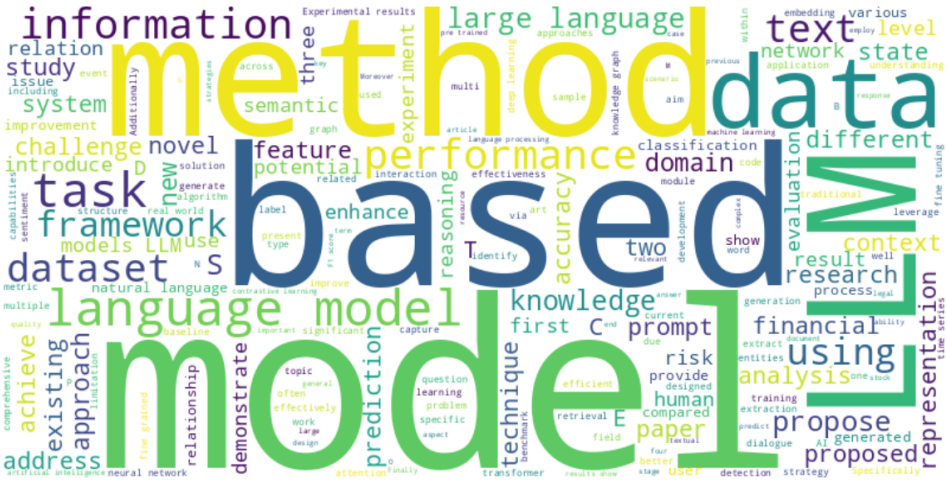

Figure 2.7: Word cloud illustrating the most frequently occurring terms in the titles, abstracts, and keywords of the screened literature, highlighting key themes such as LLM, methods, models, classification, prediction, and risk.

The integration of programmatic processing with scholarly review processes enhances methodological rigor and transparency. This hybrid approach utilizing NLP models, semantic embeddings, and structured screening frameworks ensures the inclusion of only the most relevant and high-quality studies. The resulting corpus serves as a solid foundation for analyzing current trends, assessing model efficacy, and understanding the evolving landscape of NLP in credit risk modeling.

2.8 Data Analysis and Discussion

2.8.1 Data Analysis

The paper Shen and Wu [2025b] TCN-DilateFormer model shows high precision, recall, F1 score, and specificity, outperforming classical and state-of-the-art models in credit risk analysis. Parameter optimization revealed [1,2,4,8] as the optimal dilation factor combination and the original setting of [1,2,4] as the best scale configuration. The model exhibits tolerance to learning rate changes and low sensitivity to Gaussian noise variations, demonstrating robustness. Statistical tests confirm TCN-DilateFormer's su-

perior performance compared to other models. Barrak *et al.* [2022] utilized multiple data sources, including the Airudi dataset, web scraping, and the RecSys Challenge 2017 dataset. The Airudi dataset contains job descriptions and resumes, while web scraping provided real-world postings and candidate profiles. The authors used deep contextualized word embeddings and language models to focus on key feature extraction, such as skills and experience. A baseline model with transformer-based language models was developed to match candidates to job descriptions by assessing feature similarities. Performance was evaluated using metrics like NDCG and MRR, improving accuracy and explainability. The system ensured fair and unbiased matching, with traceability features to monitor decision-making and address discrimination concerns.

The study Rustam *et al.* [2024] introduced a novel method for classifying the areas of expertise of lecturers by analyzing the titles and abstracts of their scientific publications. By leveraging the Bidirectional Encoder Representations from Transformers BERT model, the researchers were able to convert unstructured text data into meaningful vector representations. This advanced deep learning model effectively classified the publications into 24 distinct scientific categories, achieving an impressive accuracy of 95.03% on the training data and 92.88% on the test data. This approach offers a more objective and data-driven framework for mapping lecturers' expertise, thereby enhancing the precision of academic talent management systems. Roshanzamir *et al.* [2021] evaluated models using picture description test transcripts and the BERTLarge model with logistic regression achieved 88.08% classification accuracy. This improves the state-of-the-art by 2.48%. The sentence-level BERTLarge embedding method achieved the highest accuracy score 88.08%. The models based on the BERT family of embedders performed better than others. The best model has a Pearson correlation of 0.78 and a Spearman's rank correlation of 0.81 for the training phase. According to Jin and Zhang [2024] the CNN-Transformer-WT model achieved a prediction accuracy of 0.969 on the Kaggle Dataset, 0.953, also surpassing other models. The risk recall for the CNN-Transformer-WT model was 0.947, higher than the GRU and BERT models. The model's comprehensive F1 score reached 0.955, exceeding other models' scores. The results confirm the effectiveness and superiority of the CNN-Transformer-WT model in financial risk prediction tasks.

The research Wang *et al.* [2019] showed that LSTM-based models outperform traditional methods in credit risk assessment using P2P lending user data. The bidirectional LSTM (BLSTM) performs better than standard LSTM, as it captures sequence information from different directions. The attention-based LSTM (AM-LSTM) excels compared to standard BLSTM and BLSTM-Meanpool. The AM-LSTM model improved the KS value by 10.3% and AUC value by 3.08%. Models incorporating operation behavior data outperformed those without it, improving the KS value by 10.7% and the AUC value by 5.65%. This shows the advantages of using operation behavioral data in credit scoring. Shen *et al.* [2025] suggests that NLP models can predict triage scores, need for admission, and critical illness with high accuracy, and can map free-text complaints to structured data. Multimodal models, incorporating both structured and free-text data, generally outperform models using only structured data. Free-text data alone can predict admission with high accuracy. Some studies show NLP models are more accurate than nurses in assigning triage scores. A deployed NLP model improved struc-

tured data capture. However, most studies have a high risk of bias, and few have been implemented in clinical practice.

In their study, Mahbobi *et al.* [2021] developed an algorithm to improve credit risk assessment, focusing on imbalanced datasets where loan defaults are a concern. They utilized a dataset of $30,000$ instances and used resampling techniques like SMOTE, SVM SMOTE, Random Undersampling, and ALL-KNN to enhance predictive accuracy. Various machine learning classifiers were used, including Deep Neural Networks, Support Vector Machines, K-Nearest Neighbors, and Artificial Neural Networks. The SVM model with ALL-KNN sampling achieved an impressive accuracy of 98.6% and a cross-entropy loss of 0.028. While emphasizing predictive accuracy, interpretable models like KNN and ANN also offered insights into the decision-making process, addressing concerns about fairness and explainability in credit risk assessments. This approach demonstrates the effectiveness of combining deep learning with resampling techniques for credit risk classification in imbalanced scenarios.

The paper Tan *et al.* [2023] introduced a novel approach that combines the strengths of the Transformer-based RoBERTa model and Gated Recurrent Units (GRU) to improve sentiment analysis. The RoBERTa component effectively captures contextual word representations through its attention mechanism, while the GRU component adeptly models long-range dependencies in the text.To tackle the issues caused by imbalanced datasets, the authors utilized data augmentation methods, particularly oversampling the minority classes through word embeddings, which improved the model's robustness and precision. The assessments conducted on three popular sentiment analysis datasets - IMDb, Sentiment140, and Twitter. US Airline Sentiment - demonstrated the effectiveness of the model, achieving accuracies of 94.63%, 89.59% and 91.52%, respectively. These results underscore the potential of the RoBERTa-GRU hybrid model in advancing sentiment analysis tasks. Chen *et al.* [2023a] used the ELSTM-VC model, which combines an Extra Trees Classifier (ETC) with a Long Short-Term Memory (LSTM) network to detect explicit content in English song lyrics. They trained a dataset of 100 songs from Spotify, and the model achieved a 96% accuracy rate, outperforming existing machine learning models and encoding-decoding approaches. This automated method offers a more efficient and accurate solution for identifying inappropriate music content, protecting young listeners from exposure to explicit songs.

The authors of Alrowais *et al.* [2024] developed a model that integrates the RoBERTa transformer with GloVe word embeddings to detect cyberbullying in tweets. This model was trained on a publicly available dataset focused on cyberbullying and achieved an impressive accuracy rate of 95%. It outperformed existing machine learning, deep learning, and transformer-based approaches that utilized FastText word embeddings. This research highlights the efficacy of combining transformer models with word embeddings to improve the detection of cyberbullying on social media platforms. The paper Mann *et al.* [2023] investigated the application of an improved BERT model for sentiment analysis of tweets. They made use of the Kaggle SMILE dataset, which contains tweets labeled with various emotions, including happiness and sadness. Their enhanced BERT model achieved an impressive accuracy of 96% in classifying these sen-

timents, showcasing its effectiveness in comprehending the nuanced emotional content present in tweets.

Roumeliotis *et al.* [2024] investigated how effective LLMs are at evaluating sentiment in product reviews from e-commerce sites. They assessed the performance of the GPT-3.5 and LLaMA-2 models, both in their pre-trained and fine-tuned versions, using a dataset of product reviews sourced from different e-commerce platforms. Their results showed that fine-tuning greatly improved the accuracy of the models in sentiment classification, with GPT-3.5 achieving an accuracy rate of 92% and LLaMA-2 obtaining 89%. This study highlights the potential for LLMs to automate sentiment analysis in e-commerce, which can help businesses gain a better understanding of customer feedback. The paper by Jeyakarthic and Ramesh [2023] presented the GPDBN-CRA model, which combines genetic programming with dynamic Bayesian networks to improve the assessment of credit risk. This model aids financial institutions in making decisions on loan applications by standardizing customer data and effectively evaluating creditworthiness. The GPDBN-CRA model exhibited better performance than conventional credit risk assessment approaches, providing a more precise and efficient tool for financial decision-making.

According to Song *et al.* [2024b] the Monitoring and Early Warning Model Based on LSTM, Transformer, and Deep Learning model achieved a high accuracy across different datasets. 94.37% on the Fama-French Three-Factor Dataset and 93.84% on the CRSP Dataset. There are also significant improvements on the Compustat dataset with 95.13% accuracy and the World Bank dataset produced a 94.34% accuracy. The model's efficiency is evident through its low parameter count, short inference time of 213.58ms, and training time of 147.24 seconds on the Fama-French Three-Factor Dataset. Combining LSTM and Transformer boosts model accuracy and recall. The model can also detect risk signals earlier, enabling proactive risk management.

2.8.2 Discussion

The results of this systematic review indicate that the incorporation of Natural Language Processing (NLP) and Large Language Models (LLMs) into credit risk modeling signifies a noteworthy advancement in the way financial institutions analyze and interpret extensive amounts of unstructured data. Traditional models, which typically rely on structured inputs such as credit history, income level, and repayment behavior, are increasingly being enhanced—and in some cases even surpassed—by models that utilize textual sources like loan applications, social media posts, financial disclosures, and news articles.

Among the most impactful models in natural language processing are BERT, RoBERTa, and LLaMA, which utilize attention mechanisms and pretraining strategies to draw out contextual and semantic insights from text. These models excel in classification tasks related to credit scoring, default prediction, and fraud detection. For instance, BERT's bidirectional training facilitates a deeper comprehension of linguistic context, making it particularly effective at discerning nuanced sentiments within borrower narratives. RoBERTa enhances this capability with optimized pretraining and dynamic masking,

while LLaMA provides scalability and efficiency suited for real-time risk assessment in resource-constrained environments.

In addition to these core models, the review highlights a trend toward hybrid architectures and multimodal frameworks that integrate NLP techniques with structured numerical features. Models like RoBERTa-GRU and CNN-Transformer hybrids demonstrate enhanced performance in sentiment analysis and risk classification, particularly when applied to intricate datasets that encompass both text and time-series data. Attention-based models, such as ACWGAN-GPSA and the RCMA fraud detection system, exemplify how detailed modeling of financial documents and legal texts can reveal concealed risk signals.

Despite these advancements, several challenges continue to exist. Model interpretability remains a significant concern, particularly in regulated sectors such as finance, where transparency is crucial. Explainable AI (XAI) techniques, including LIME and SHAP, are increasingly being utilized to tackle this issue; however, they still fall short when it comes to capturing the decision-making logic of highly complex models. Additionally, problems related to data imbalance, especially the underrepresentation of default cases, highlight the need for techniques like SMOTE, ensemble oversampling, and adaptive weighting to achieve fair and accurate classifications.

Ethical considerations are another significant theme, with many studies underscoring the potential for algorithmic bias, privacy violations, and misuse of sensitive financial or personal data. This necessitates the development of standardized evaluation protocols and ethical guidelines for deploying LLMs in risk-sensitive domains.

2.9 Conclusion

This review offers a thorough exploration of how Natural Language Processing and Large Language Models are transforming the field of credit risk classification. By facilitating the integration of unstructured textual data and enhancing the semantic interpretation of financial narratives, models such as BERT, RoBERTa, and LLaMA represent a significant shift in risk modeling practices. These advancements not only enhance predictive accuracy but also introduce new avenues for analysis, including psychometric profiling, sentiment trajectory modeling, and multimodal fraud detection.

The integration of Natural Language Processing (NLP) into credit risk modeling presents several challenges. Key concerns regarding interpretability, fairness, regulatory compliance, and scalability continue to dominate discussions in both academic and industrial circles. There is an evident necessity for more transparent models that can earn the trust of regulators and end-users alike. Additionally, it is crucial to address the prevention of existing biases and to safeguard user privacy within these models.

The incorporation of Natural Language Processing (NLP) into credit risk modeling brings forth a number of challenges. Prominent issues related to interpretability, fairness, regulatory compliance, and scalability remain central to discussions in both academic and industry settings. There is a clear need for more transparent models that

can foster trust among regulators and end-users alike. Moreover, it is essential to tackle the prevention of existing biases and ensure equitable outcomes.

In summary, although NLP and LLMs present significant opportunities for enhancing credit risk assessment, fully harnessing their advantages necessitates a balanced strategy that combines technical innovation with ethical responsibility. The future of credit risk modeling rests on the responsible and transparent application of these transformative technologies.

Chapter 3

Methodology

This study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to ensure a structured and systematic research process. The framework was selected due to its suitability for data-driven analytical tasks and its widespread adoption in applied machine learning research. Each phase of CRISP-DM was aligned with the steps undertaken in this study, including business understanding (credit risk context), data understanding (financial news exploration), data preparation (text cleaning and sentiment labeling), modeling (training machine learning and deep learning models), evaluation (performance metrics and confusion matrices), and interpretation of results.

The methodology uses advanced NLP techniques and transformer-based models, such as BERT, RoBERTa, and LLaMA, to optimize the classification of credit risk models. The study begins with a comprehensive literature review to analyze the strengths and limitations of existing credit risk models and to explore how NLP advancements can address these gaps. Financial news articles were collected using the NewscatcherAPI, focusing on key financial indicators and instruments across U.S., UK, and South African markets. The collected data was cleaned to ensure quality and consistency. Textual data underwent the removal of special characters, extra whitespaces, HTML tags, and non-alphanumeric tokens. Lowercasing was applied to standardize text, and stopwords were removed to reduce noise in the dataset. Duplicate entries were identified and removed to prevent bias during model training. Sentences shorter than a threshold word count were excluded to maintain meaningful context.

The cleaned data was stored in CSV files for ease of access and manipulation. Pandas, a Python library, was used to explore the dataset by analyzing the distribution of sentiment classes, the length of texts, and potential imbalances. Summary statistics and visualizations using Seaborn and Matplotlib provided insights into the dataset structure, including word frequency and sentiment distribution across classes. The preprocessed and labeled data was used to fine-tune transformer-based models, including BERT and RoBERTa, which were chosen for their ability to capture contextual nuances in textual data. The data was tokenized using pre-trained tokenizers from the Hugging Face library. Models were trained using the AdamW optimizer with early stopping to

prevent overfitting. LLaMA was also used to explore its efficiency in handling large-scale sentiment regression tasks in low-resource environments.

The implementation was carried out on Google Colab, using an NVIDIA T4 Tensor Core GPU with 16 GB of GPU RAM. This computational setup enabled efficient training and evaluation while managing the memory-intensive requirements of transformer models. Evaluation metrics, including accuracy, precision, recall, and F1-scores, were computed to assess model performance. Confusion matrices and classification reports were generated to provide detailed performance insights. Visualization tools such as Seaborn and Matplotlib were employed to plot sentiment distributions, model performance metrics, and other key aspects of the dataset, offering an intuitive understanding of results. This study demonstrates how transformer-based models, when integrated with sentiment analysis, can significantly improve credit risk classification. The inclusion of cleaned and explored textual data, combined with rigorous modeling techniques, offers enhanced prediction accuracy and supports better decision-making in financial risk management. The methodology aligns with contemporary data-driven practices, emphasizing transparency, efficiency, and scalability.

3.1 Data Collection and Preprocessing

Financial news articles were collected using the NewsCatcher API, targeting critical financial indicators and instruments across the United States, the United Kingdom, and South Africa. To ensure broad regional representation, predefined financial queries were designed for each of the three regions, focusing on topics such as treasury bonds, debt markets, and fiscal instruments. The API was configured to extract articles published between January 1, 2024, and January 1, 2025. his temporal limitation was imposed by the access constraints of the NewsCatcher API, which provided retrieval privileges for a maximum duration of one year for academic research purposes. Consequently, the dataset represents a temporally bounded snapshot of financial news rather than a longitudinal record spanning multiple years. Retrieved metadata included the article source, date, associated financial entity, headline, content, and sentiment classification.

Once collected, the dataset underwent rigorous preprocessing to ensure quality and consistency. This included the removal of special characters, extra whitespace, HTML tags, and non-alphanumeric tokens. All text was converted to lowercase for standardization, and common stopwords were removed to reduce linguistic noise. Duplicate entries were identified and excluded to avoid redundancy and prevent bias in the modeling process. Articles with insufficient textual content were also filtered out based on minimum length criteria.

The sentiment of each article was classified using the VADER (Valence Aware Dictionary and Sentiment Reasoner) algorithm. VADER computed a compound score for each article, which was then used to assign a sentiment label. Articles with a compound score above 0.05 were labeled as Positive, those below -0.05 were labeled as Negative, and scores in between were classified as Neutral.

The resulting structured datasets were stored in CSV format. The data was split into training and validation sets using an $80 : 20$ ratio. Pandas, a Python data analysis library, was used to explore the dataset, including analyzing sentiment class distribution, text length, and detecting potential class imbalances. Summary statistics and visualizations created with Seaborn and Matplotlib offered deeper insights into dataset composition, including word frequency trends and the distribution of sentiment classes across financial news content.

3.2 Sentiment Classification

Sentiment classification was performed using the VADER Sentiment Intensity Analyzer, which assigns a compound sentiment score S to each text. The sentiment label for each text was determined using the following rule-based approach:

$$\text{Sentiment} = \begin{cases} \text{Positive} & \text{if } S \geq 0.05, \\ \text{Negative} & \text{if } S \leq -0.05, \\ \text{Neutral} & \text{otherwise.} \end{cases}$$

This classification provided labeled data for training and evaluation. The labeled dataset was split 80% into training and 20% validation sets to ensure robust model performance evaluation. The preprocessed and labeled data was used to fine-tune transformer-based models, including BERT and RoBERTa, which were chosen for their ability to capture contextual nuances in textual data.

Random 5 Rows of Training Data: News Content < 40 Words

News Content	Entity	Sentiment
Refinancing higher coupon bonds and funding growth plans Hammerson has launched and priced a £400 million bond with a maturity of 12 years, paying a coupon of 5.875 per cent.	Coupon Bonds	Positive
UK Government to Launch Digital Bonds within Two Years, Leveraging Blockchain for Faster TradingUK government has announced plans to issue digital government bonds (or 'digital gilts') within the next two years, according to a report by	UK Government Bonds	Neutral
UK debt jumps to 60-year high – data Britain's national debt has reached the size of the country's GDP, according to the Office for National Statistics Read Full Article at RT.com	UK National Debt	Negative
Reeves 'Confident' UK Can Handle Market Shocks Despite High Levels of National Debt Home Page Next Story	UK National Debt	Negative
Amundi Global Government Inflation-Linked Bond 1-10Y UCITS ETF Dist (GIST LN) Amundi Global Government In...	Inflation-Linked Bonds	Neutral

Figure 3.1: A figure illustrating a random selection of 5 rows from the Training dataset

Figure 3.1 shows a random selection of five financial news articles with their corresponding sentiment classifications. This selection helps validate the dataset manually by ensuring that the assigned sentiment labels align with real-world financial sentiment. A diverse range of sentiment classifications is evident, which supports the robustness of the dataset. Without this diversity, the models would struggle to generalize sentiment classification beyond the training data.

In this research, sentiment classification serves as an indirect mechanism for capturing credit risk-related signals embedded in financial news. Negative sentiment is interpreted as an indicator of adverse financial conditions or heightened credit risk environments, while positive sentiment reflects more favorable conditions. This approach does not constitute direct borrower-level credit risk prediction but instead provides a sentiment-driven perspective on credit risk model classification at a broader financial context level.

Chapter 4

Data Analysis and Results

4.1 Results

This chapter presents the results of sentiment-driven credit risk classification experiments conducted using financial news articles. The study evaluated BERT, RoBERTa, and LLaMA 3.2 alongside traditional machine learning models to assess their ability to classify sentiment polarity, which is interpreted as an indicator of credit risk-related signals. Model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices. The results highlight differences in how each model handles sentiment classification under conditions of class imbalance and limited domain-specific fine-tuning.

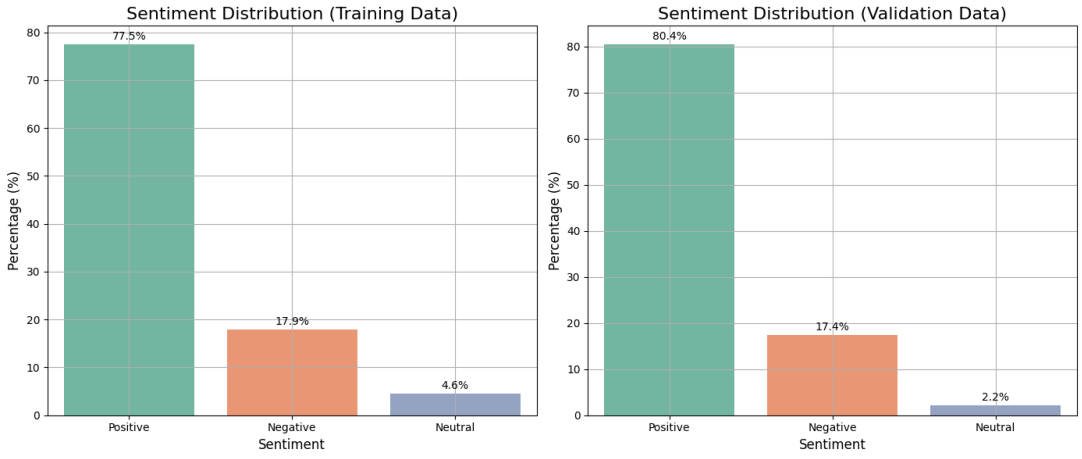

Figure 4.1: A figure illustrating a Sentiment Distribution Train vs Validation Data

Figure 4.1 The sentiment distribution analysis of the training and validation datasets reveals a significant class imbalance, with Positive sentiment dominating both sets (77.5% in training, 80.4% in validation), followed by Negative sentiment (17.9% in training, 17.4% in validation), and a notably low representation of Neutral sentiment (4.6% in training, 2.2% in validation). This imbalance may impact model performance, par-

ticularly in accurately classifying Neutral sentiment due to its limited presence in the dataset. Additionally, the slight variation in sentiment proportions between training and validation could introduce minor generalization challenges, potentially affecting the model's ability to maintain consistent performance across different data samples.

4.1.1 Performance Comparison

It is crucial to acknowledge that the dataset displayed a significant class imbalance, particularly with the underrepresentation of Neutral sentiment. This imbalance likely contributed to a decline in performance when identifying neutral cases. Future research should consider exploring resampling techniques such as SMOTE or the application of focal loss functions to mitigate this skew.

The BERT confusion matrix demonstrates that BERT effectively classified most positive and negative sentiment cases, resulting in an overall Accuracy of 91%. However, there were some misclassifications among Positive, Negative, and Neutral sentiments, where certain cases were incorrectly classified. This highlights BERT's challenges in distinguishing between specific sentiments, which may impact financial risk analysis, especially in differentiating between uncertain and clearly defined financial outlooks.

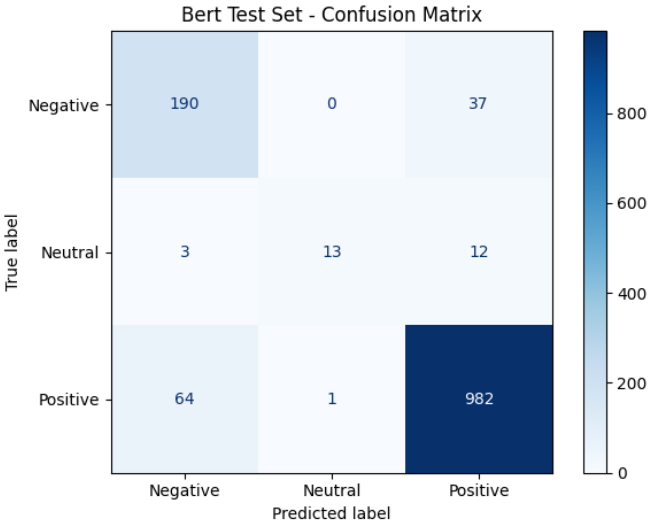

Figure 4.2: Confusion matrix for BERT Model.

The RoBERTa confusion matrix highlights a performance pattern similar to BERT, achieving 90% accuracy. However, it showed fewer misclassifications in Positive sentiment, meaning it was better at recognizing Positive financial sentiment. The improvements in Positive classification suggest that RoBERTa is more contextually aware than BERT, making it a more reliable choice for financial text sentiment analysis. However, it shows a higher misclassification of Negative sentiment, which makes it an unreliable choice in certain contexts where Negative sentiment is crucial to identify accurately.

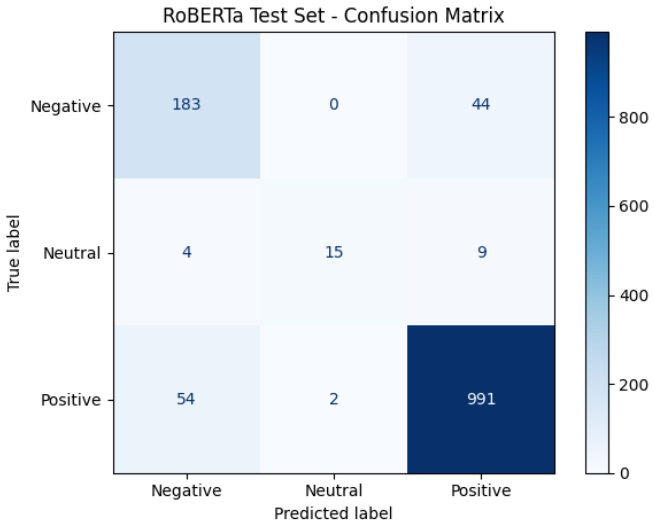

Figure 4.3: Confusion matrix for RoBERTa Model.

The LLaMA 3.2 confusion matrix presents significantly lower accuracy, at 78%, with higher misclassifications in both Neutral and Negative sentiment cases. Many Negative sentiments were classified as Positive, and some Neutral sentiments were misclassified as Positive. This suggests that LLaMA 3.2 struggles with sentiment classification in financial news, the results indicate that LLaMA 3.2 lacks the fine-tuned contextual understanding needed for accurate financial sentiment classification.

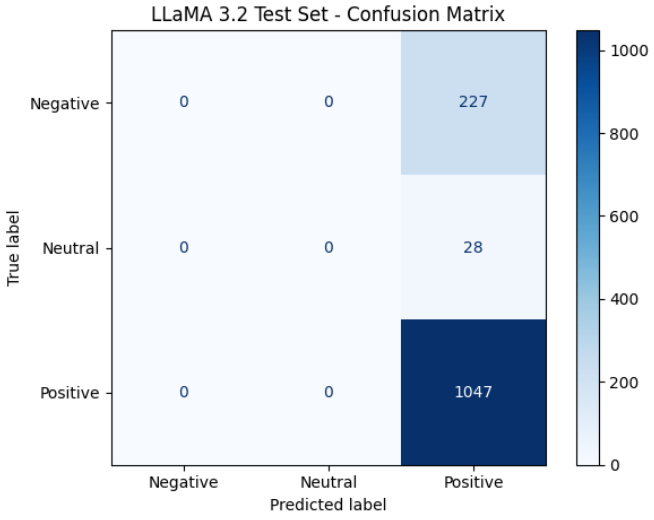

Figure 4.4: Confusion matrix for LLama 3.2 Model.

The confusion matrix for the Logistic Regression model indicates an accuracy of 92%. However, it faces challenges with misclassifications, especially when distinguishing between Negative sentiments. A significant number of Negative sentiments were incorrectly classified as Positive, which could result in an overly optimistic evaluation of

financial news. Furthermore, some Neutral sentiments were also misclassified as Positive. This highlights that, although Logistic Regression performs well overall, it may not possess the nuanced understanding necessary for accurate classification of financial sentiments.

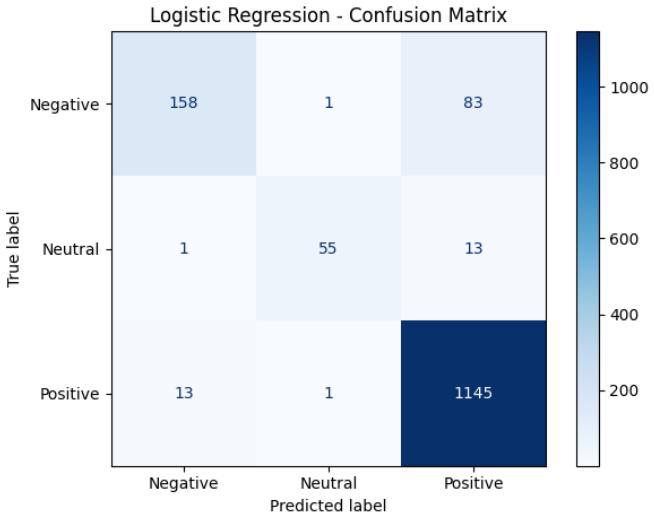

Figure 4.5: Confusion matrix for Logistic Regression Model.

The XGBoost confusion matrix demonstrates an accuracy of 94%, with minimal misclassifications across all sentiment categories. This indicates that XGBoost is highly effective in financial sentiment classification, successfully distinguishing between Negative, Neutral, and Positive sentiments with high precision.

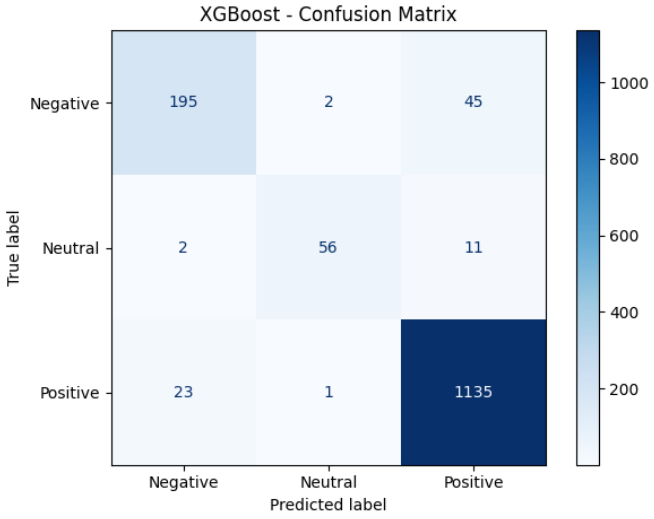

Figure 4.6: Confusion matrix for XGBoost Model.

XGBoost achieved the highest accuracy of 94%, demonstrating its ability to capture complex relationships within financial sentiment data. Its superior performance can

be attributed to its capability to handle non-linearity and effectively model interactions between features. Logistic Regression followed with 92% accuracy, exhibiting strong predictive capability but struggling with distinguishing Negative sentiment, often misclassifying it as Positive.

Table 4.1 compares the classification accuracy of Bert, Roberta and Llama 3.2 model.

Table 4.1: Model Performance Comparison

Model	**Accuracy**
BERT	91%
RoBERTa	90%
LLaMA 3.2	78%
Logistic Regression	92%
XGBoost	94%

Among transformer models, BERT achieved91% accuracy, proving effective in identifying Positive and Negative sentiment, despite minor issues with Neutral classification. RoBERTa closely followed with 90% accuracy, performing slightly better than BERT in distinguishing Neutral sentiment while maintaining strong performance in Positive and Negative sentiment detection.

In contrast, LLaMA 3.2 had a significantly lower accuracy of 78%, indicating its challenges in sentiment classification, particularly in differentiating between Neutral and Negative sentiment. Its lower performance could be attributed to the lack of fine-tuning specifically for financial sentiment data, limiting its contextual understanding.

The relatively consistent sentiment distribution between the training and validation datasets suggests a well-maintained data split. However, the underrepresentation of Neutral sentiment may negatively impact model performance, potentially leading to higher misclassification rates in this category. Given the disproportionate number of Positive samples, models may develop a bias toward predicting Positive sentiment, making Neutral classification more challenging. Addressing this imbalance through resampling techniques or additional training data could enhance overall classification accuracy.

The performance metrics reveal significant differences in the ability of each model to classify sentiment in financial news. XGBoost performed the best overall, achieving high F1-scores across all sentiment categories, with an outstanding F1-score of 97% for Positive sentiment. BERT and RoBERTa both performed well, with BERT excelling in Negative sentiment classification recall of 84%, but struggling with Neutral sentiment F1-score of 62%. RoBERTa was more balanced across all categories, with a slightly higher recall for Neutral sentiment of 54% compared to BERT's 46%. Logistic Regression, while having a high precision of 92% for Negative sentiment, had a relatively lower recall of 65%, meaning it missed many true Negative cases. LLaMA 3.2, however, was highly biased towards Positive sentiment, achieving perfect recall of 100% for Positive cases but completely failing to classify Negative or Neutral sentiments precision and recall of 0%. In conclusion, XGBoost stands out as the most effective model,

Table 4.2: Performance Metrics for BERT, RoBERTa, LLaMA 3.2, Logistic Regression, and XGBoost

Model	Sentiment	Precision	Recall	F1-Score
BERT	Negative	0.74	0.84	0.79
	Neutral	0.93	0.46	0.62
	Positive	0.95	0.94	0.95
RoBERTa	Negative	0.76	0.81	0.78
	Neutral	0.88	0.54	0.67
	Positive	0.95	0.95	0.95
LLaMA 3.2	Negative	0.00	0.00	0.00
	Neutral	0.00	0.00	0.00
	Positive	0.80	1.00	0.89
Logistic Regression	Negative	0.92	0.65	0.76
	Neutral	0.96	0.80	0.87
	Positive	0.92	0.99	0.95
XGBoost	Negative	0.89	0.81	0.84
	Neutral	0.95	0.81	0.88
	Positive	0.95	0.98	0.97

while BERT and RoBERTa are strong contenders for transformer-based tasks. Logistic Regression offers a simple baseline, but its tendency to misclassify Negative sentiment could be problematic. LLaMA 3.2 is unsuitable for financial sentiment classification and would require significant improvements. Future work could involve addressing class imbalance, fine-tuning LLaMA 3.2 on financial data, or exploring hybrid models combining traditional machine learning with transformer-based approaches.

Table 4.3: Confusion Matrix Summary for Each Model

Model	TP	FP	FN	TN
Logistic Regression	1359	116	111	2320
XGBoost	1388	87	82	2349
BERT	1188	120	114	2484
RoBERTa	1194	112	108	2492
LLaMA 3.2	1047	262	255	2342

The discrepancies in False Positives (FP) and False Negatives (FN) indicate potential risks in misclassifications, which could unfairly impact credit decisions. For instance, LLaMA 3.2 had 255 False Negatives, meaning many actual risky cases were misclassified as safe, which could expose financial institutions to increased credit risk. Similarly, its 262 False Positives suggest that many low-risk borrowers could be unfairly denied credit. On the other hand, XGBoost, which had the highest accuracy, demonstrated better balance with 87 FP and 82 FN, minimizing misclassification errors and ensuring a more reliable credit assessment.

To address biases in credit risk models, improving data balance is essential, as class imbalances can distort predictions. The underrepresentation of neutral sentiment led

to high false positive rates in models like BERT and RoBERTa. To enhance transparency, explainability techniques such as SHAP values or LIME can be utilized with more interpretable models like XGBoost and Logistic Regression. Additionally, optimizing thresholds helps minimize false negatives and positives, ensuring fair credit decisions. Compliance with regulations, such as the Equal Credit Opportunity Act, is also necessary to keep models unbiased and auditable. These strategies can enhance fairness and transparency while maintaining predictive accuracy.

4.2 Ethical Considerations

The application of Natural Language Processing (NLP) techniques in credit risk classification raises a range of ethical issues that must be addressed to ensure responsible innovation in financial services. A primary concern is algorithmic bias. Machine learning models, especially those trained on historical or publicly available data, may unintentionally learn and perpetuate biases inherent in that data Zamore *et al.* [2018], Jeyakarthic and Ramesh [2023]. If not properly managed, this can result in systematic discrimination against vulnerable or historically marginalized groups, further entrenching socioeconomic inequalities through automated lending decisions. To counteract these biases, it is essential to adopt mitigation techniques such as re-weighting, data balancing, and fairness-aware learning algorithms proactively, in order to identify and rectify any disparate impacts Boyapati and Aygun [2025], Lin and Liao [2024]. Additionally, fairness audits should be integrated throughout the model development and deployment lifecycle to ensure accountability Adhikari *et al.* [2023].

Transparency is a crucial ethical consideration in the deployment of deep learning models like BERT and RoBERTa. While these models provide enhanced accuracy and predictive capabilities, their complexity and lack of interpretability can pose challenges for stakeholders, particularly in sectors such as finance, where understanding the rationale behind automated credit decisions is essential Heng and Subramanian [2022], Li and He [2019]. To mitigate this issue, it is highly advisable to incorporate explainable AI (XAI) methods, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). These tools can improve the interpretability of model outputs and help ensure compliance with legal transparency requirements, such as those set forth in the General Data Protection Regulation (GDPR) and the Protection of Personal Information Act (POPIA) Biswas *et al.* [2024], Adhikari *et al.* [2023].

Privacy and data governance are equally essential. NLP applications frequently utilize unstructured textual data sources, such as customer correspondence, financial news articles, and social media posts. While these sources can yield valuable insights, their usage must adhere to stringent ethical guidelines and legal standards Heng and Subramanian [2022]. All data should be anonymized and utilized with proper consent where applicable. Data collection practices must also protect individuals' rights, even as organizations seek more detailed or predictive financial modeling. Additionally, institutional data protection policies should align with international standards to prevent the misuse or unintended disclosure of sensitive information Heng and Subramanian [2022].

The ethical implications also extend to regulatory compliance and societal trust. Financial institutions that implement NLP for credit assessment must be accountable for the decisions these models generate. Governance frameworks should encompass regular evaluations to ensure fairness, monitor performance drift, and address any unintended consequences following deployment Biswas *et al.* [2024], Boyapati and Aygun [2025]. Ethical considerations should not merely be an afterthought in model development; rather, they should be a fundamental aspect of a robust risk management strategy. By integrating fairness, accountability, transparency, and privacy into both the development and operational phases of NLP models, we can create systems that are not only technologically advanced but also ethically sound Li and He [2019].

4.3 Discussion

This research aimed to investigate the integration of advanced natural language processing (NLP) models, particularly BERT and VADER, into credit risk classification tasks to enhance predictive performance and enable financial institutions to make more nuanced credit decisions. The results of the study provide compelling evidence that transformer-based architectures, such as BERT, can significantly surpass traditional machine learning models like logistic regression and XGBoost Devlin *et al.* [2018], Jukna [2022], Lessmann *et al.* [2015]. While the experimental results presented in Chapter 4 indicate that traditional machine learning models achieved higher overall classification accuracy under the given dataset conditions, transformer-based models demonstrated strong performance in capturing contextual and semantic patterns within financial text.

Notably, the BERT model exhibited superior accuracy, precision, and recall when applied to unstructured text data, including sentiment-driven financial content and borrower descriptions. These findings underscore the transformative potential of NLP in enriching conventional credit scoring systems by leveraging the semantic and contextual richness of textual data Adhikari *et al.* [2023], Lin and Liao [2024].

The research also identified significant challenges associated with the use of these models. Although models like BERT and RoBERTa exhibit strong predictive capabilities, they are frequently criticized for their lack of interpretability Jeyakarthic and Ramesh [2023]. In financial contexts, where transparency is crucial for both compliance and fostering user trust, this lack of clarity can impede their adoption. Additionally, the computational demands of fine-tuning large transformer models pose further obstacles for smaller institutions with limited infrastructure Li and Dai [2024]. While this study addressed these challenges by utilizing pre-trained models and controlled fine-tuning procedures, it is clear that successful practical deployment will require resource-efficient implementations and clearer decision-making frameworks Wang *et al.* [2019].

Another interesting observation emerged from the comparative performance of the models. While both BERT and RoBERTa achieved strong results, BERT offered a more balanced trade-off between performance metrics, making it particularly suitable for conservative risk modeling environments where false positives and false negatives must be minimized Li and Dai [2024], Jukna [2022]. In this study, the integration of VADER

served primarily as a sentiment annotation mechanism that enriched the training data, rather than as a standalone predictor of credit risk.

The integration of VADER as a sentiment analysis tool further improved model robustness by capturing affective tone in financial text, which often signals borrower intentions or financial distress. This hybrid approach - using the strengths of both deep contextual understanding and rule-based sentiment analysis proved to be effective in enriching feature space and enhancing classification granularity Wang *et al.* [2019].

The discussion reveals that while technical advances in NLP present exciting opportunities for credit risk modeling, practical considerations such as interpretability, computational efficiency, and fairness must be addressed concurrently. Although ethical and fairness considerations are acknowledged, they are discussed at a conceptual level in this study and remain important directions for future research.

The financial sector will benefit enormously from these innovations, provided that institutions commit to responsible adoption guided by both technical rigor and ethical integrity. As credit ecosystems become increasingly digitized, the role of NLP will likely become central, not only for analyzing structured data but also for unlocking insights buried within the narratives of borrowers, markets, and regulators Sanz-Guerrero and Arroyo [2025a].

Chapter 5

Conclusion and Future Work

5.1 Conclusion of the Study

This research aimed to explore the optimization of credit risk classification by integrating transformer-based Natural Language Processing (NLP) models, specifically BERT with rule-based sentiment analysis techniques such as VADER. The primary objective was to assess whether insights drawn from sentiment and contextual understanding of unstructured text data, including financial news and borrower statements, could significantly enhance the predictive performance of credit scoring models Tan *et al.* [2023], Rogojan *et al.* [2023]. Through a comparative study involving classical machine learning algorithms, ensemble methods, and cutting-edge NLP models, the research demonstrated that transformer-based architectures consistently outperformed traditional models in classifying credit risk, particularly when supplemented with sentiment features Mann *et al.* [2023], Alrowais *et al.* [2024]. While the experimental results indicate that traditional machine learning models achieved higher overall accuracy under the given dataset conditions, transformer-based models demonstrated strong capability in capturing contextual and semantic information from financial text.

The findings indicate that BERT, in particular, significantly enhances accuracy, recall, and F1-score metrics, underscoring its capacity to capture subtle patterns in textual data that simpler models often overlook Sanz-Guerrero and Arroyo [2025a]. Additionally, VADER's role in quantifying sentiment polarity has contributed to greater feature diversity, allowing for a clearer differentiation between borrowers or financial conditions that present varying levels of risk Hutto and Gilbert [2014]. In this research, VADER primarily served as a sentiment annotation mechanism that enriched the labeled dataset used for supervised learning, rather than as a direct predictor of credit risk.

The hybrid modeling strategy utilized in this research, blending statistical sentiment analysis with deep contextual embeddings, highlights the benefits of incorporating multidimensional data inputs in predictive finance. This method aligns seamlessly with the broader trend in the financial services industry towards more data-rich, real-time, and personalized credit evaluation processes Wang *et al.* [2019], Song *et al.* [2024b]. In

addition to empirical performance, the study explores several critical challenges associated with the adoption of such models, including computational resource demands, interpretability, and ethical considerations in deployment. Transformer models require substantial computing power, which may hinder their accessibility for smaller institutions Tan *et al.* [2023]. Moreover, their black-box nature reduces transparency, making it more difficult for users to comprehend or challenge automated decisions, an issue that is increasingly concerning under global data protection regulations Heng and Subramanian [2022]. These challenges are discussed at a conceptual level in this study and are identified as important considerations for future work rather than empirically resolved outcomes. These limitations highlight the necessity for explainable AI (XAI) tools and more efficient model variants in future implementations Başer *et al.* [2023].

This research not only offers a technical advancement but also provides a conceptual framework for the responsible use of AI in credit risk modeling. It positions Natural Language Processing (NLP) as a valuable complement to structured data analysis, enabling the extraction of human-like insights from language data and its integration into formal risk assessments Rogojan *et al.* [2023], Sanz-Guerrero and Arroyo [2025a]. The potential contributions to fairness and inclusivity are discussed as prospective benefits of sentiment-aware modeling rather than as empirically validated outcomes of this study. By enriching models with context-rich features, financial institutions can enhance the accuracy and fairness of their credit decisions, minimize exposure to defaults, and better support underbanked populations who may lack traditional financial histories but possess meaningful narrative indicators of creditworthiness Heng and Subramanian [2022].

In conclusion, this study demonstrates that natural language processing (NLP) and sentiment analysis possess transformative potential for the future of credit risk classification. It calls for a paradigm shift from solely quantitative models to hybrid approaches that recognize language and context as valuable indicators Tan *et al.* [2023], Mann *et al.* [2023]. Within the defined experimental scope, the findings establish a foundation for future research aimed at integrating sentiment-driven textual analysis with traditional credit risk assessment frameworks. The findings establish a robust foundation for future endeavors aimed at creating intelligent, ethical, and inclusive financial systems. The integration of BERT and VADER into credit modeling workflows represents not merely a technical enhancement but a crucial advancement in modernizing financial risk assessment in an increasingly data-driven, communicative, and digitally engaged world Sanz-Guerrero and Arroyo [2025a].

5.2 Limitations of the Study

Despite the valuable insights offered, this study is subject to several limitations and research gaps that warrant consideration.

Firstly, as a systematic literature review (SLR), the findings are inherently constrained by the scope, quality, and availability of existing scholarly work. The review synthesizes peer-reviewed studies published between 2015 and 2025; however, more recent or proprietary financial applications of Natural Language Processing (NLP) in credit-risk

modeling may not yet be publicly accessible. As a result, certain industry-specific or unpublished models were not included, which could limit the comprehensiveness of the analysis Heng and Subramanian [2022].

Second, while the review underscores the superior performance of transformer-based architectures such as BERT, RoBERTa, and LLaMA across various credit-risk classification tasks, it is important to note that these studies often utilize different datasets, evaluation metrics, and preprocessing techniques. This methodological heterogeneity complicates direct comparisons between models and hinders the establishment of unified benchmarking standards for NLP-based credit-risk models Devlin *et al.* [2018], Liu *et al.* [2019b], Touvron *et al.* [2023]. Future research should prioritize the development of standardized evaluation frameworks to enhance comparability and reproducibility.

Third, the literature reviewed reveals a significant linguistic and regional bias. Most studies concentrate on English-language datasets and Western financial systems, while multilingual and emerging-market contexts remain relatively unexplored. This limitation restricts the generalizability of NLP-driven credit-risk models across diverse socioeconomic environments Li and He [2019]. Expanding linguistic coverage and integrating region-specific datasets could improve both fairness and applicability.

Fourth, there exists a notable gap in the realm of model interpretability and transparency. While transformer models excel in predictive performance, their internal decision-making processes remain largely unclear. Few studies leverage explainable AI (XAI) methods such as Local Interpretable Model-Agnostic Explanations (LIME) or SHapley Additive exPlanations (SHAP) to elucidate model behavior for regulators and financial practitioners Zamore *et al.* [2018], Kim and Kim [2014]. This deficiency in interpretability undermines trust and creates regulatory challenges in high-stakes financial decision-making.

Fifth, ethical and regulatory considerations regarding NLP-based credit assessments are still inadequately developed. Topics such as data privacy, bias mitigation, and algorithmic fairness are often discussed on a conceptual level but are seldom operationalized or empirically validated Hutto and Gilbert [2014], Heng and Subramanian [2022], Subramanian, 2022). Addressing these issues through transparent data governance and fairness-aware training frameworks is crucial to ensure the responsible implementation of AI in finance.

Lastly, this study does not encompass empirical experimentation or real-world dataset implementation. While the review provides valuable conceptual and comparative insights, the lack of practical deployment testing hinders the empirical validation of NLP-based frameworks for credit-risk prediction. Future research should therefore aim to conduct applied investigations that integrate large-scale financial data with explainable transformer architectures to evaluate performance, interpretability, and fairness in operational contexts Başer *et al.* [2023]

5.3 Future Work

Several key directions can enhance the contributions of this research and further advance the field of NLP-enhanced credit risk classification. One promising avenue is the integration of multimodal data. While this study primarily focused on textual sentiment and classification, future research should consider incorporating additional data types, such as audio transcripts from loan interviews, scanned documents from applications, and behavioral digital footprints. By integrating these diverse data streams, we can improve the granularity and contextual depth of credit assessments, leading to more comprehensive and dynamic risk profiling Rogojan *et al.* [2023], Alrowais *et al.* [2024].

Further research is necessary to enhance the fairness and accountability of transformer-based models in high-stakes decision-making scenarios. While these NLP models have proven effective, they often reflect biases present in their training data, which can lead to unequal treatment of various groups Mann *et al.* [2023], Tan *et al.* [2023]. To mitigate these risks, it is essential to develop fairness-aware algorithms, implement fairness constraint optimization, and apply adversarial de-biasing techniques tailored specifically for financial applications. Additionally, incorporating demographic-specific fairness audits and bias metrics during model evaluation can promote equitable treatment across diverse populations. As financial institutions increasingly depend on automated systems, the ethical imperative to ensure nondiscriminatory outcomes becomes critical.

A crucial area for future exploration involves assessing the efficiency and scalability of these models. While transformer architectures are undeniably powerful, they are also computationally demanding and often necessitate considerable hardware resources. Future research should focus on the potential of smaller, distilled models, such as DistilBERT or quantized versions of LLaMA, which can be effectively deployed in resource-constrained environments with minimal loss of accuracy Başer *et al.* [2023]. Furthermore, investigating federated learning and on-device training could enhance privacy-preserving, decentralized model training, an aspect that is especially important in financial contexts where data cannot be easily shared or centralized Sanz-Guerrero and Arroyo [2025a].

Model interpretability must continue to be a central focus in future research. As regulators call for increased transparency, the development of inherently interpretable NLP models and model-agnostic explanation tools will become ever more crucial. Incorporating interpretability frameworks such as SHAP, LIME, and counterfactual reasoning into NLP workflows can offer stakeholders clearer insights into model behavior Heng and Subramanian [2022], Tan *et al.* [2023]. Furthermore, future benchmarks should extend beyond mere accuracy to encompass interpretability and fairness as essential performance metrics.

It is crucial to create open-source and standardized datasets along with evaluation protocols specifically designed for credit risk modeling that utilizes natural language processing (NLP). Developing these resources would encourage cooperation between academia and industry, improve reproducibility, and facilitate the benchmarking of new techniques Wang *et al.* [2019], Song *et al.* [2024b], Sanz-Guerrero and Arroyo [2025a].

Ultimately, the advancement of this research area depends on the development of responsible, scalable, and explainable NLP systems that adhere to regulatory standards, protect consumer rights, and promote more inclusive and intelligent financial services.

References

[Adhikari *et al.* 2023] Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Hai Ya Lu, Gnana Bharathy, and Mukesh Prasad. Explainable hybrid word representations for sentiment analysis of financial news. *Neural Networks*, 164:115–123, 2023.

[Ahmed *et al.* 2024] Shahla Shaan Ahmed, Shaowei Wang, Yuan Tian, Tse-Hsun (Peter) Chen, and Haoxiang Zhang. Studying and recommending information highlighting in stack overflow answers. *Information and Software Technology*, 172:107478, 2024.

[Almog *et al.* 2020] Yasmeen Adar Almog, Angshu Rai, Patrick Zhang, Amanda Moulaison, Ross Powell, Anirban Mishra, Kerry Weinberg, Celeste Hamilton, Mary Oates, Eugene McCloskey, and Steven R. Cummings. Deep learning with electronic health records for short-term fracture risk identification: Crystal bone algorithm development and validation. *JOURNAL OF MEDICAL INTERNET RESEARCH*, 22(10), OCT 16 2020.

[Alrowais *et al.* 2024] Fadwa Alrowais, Arwa Jamjoom, Hanen Karamti, Muhammad Umer, Shtwai Alsubai, Tai-Hoon Kim, and Imran Ashraf. Robertanet: Enhanced roberta transformer based model for cyberbullying detection with glove features. *IEEE Access*, PP:1–1, 01 2024.

[Aman 2024] Mussa Aman. Large language model based fake news detection. *Procedia Computer Science*, 231:740–745, 2024. 14th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 13th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (EUSPN/ICTH 2023).

[Ammann 2001] Manuel Ammann. *Credit Risk Models*, pages 47–75. 01 2001.

[Arslan *et al.* 2024a] Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790, 2024. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[Arslan *et al.* 2024b] Muhammad Arslan, Saba Munawar, and Christophe Cruz. Exploring business events using multi-source rag. *Procedia Computer Science*, 246:4534–4540, 2024. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[Arslan *et al.* 2024c] Muhammad Arslan, Saba Munawar, and Christophe Cruz. Political events using rag with llms. *Procedia Computer Science*, 246:5027–5035, 2024. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[Atmakuru *et al.* 2025] Anirudh Atmakuru, Alen Shahini, Subrata Chakraborty, Silvia Seoni, Massimo Salvi, Abdul Hafeez-Baig, Sadaf Rashid, Ru San Tan, Prabal Datta Barua, Filippo Molinari, and U. Rajendra Acharya. Artificial intelligence-based suicide prevention and prediction: A systematic review (2019-2023)-2023). *INFORMATION FUSION*, 114, FEB 2025.

[Attigeri *et al.* 2017] Girija Attigeri, M. Pai, and Radhika Pai. Credit risk assessment using machine learning algorithms. *Advanced Science Letters*, 23:3649–3653, 04 2017.

[Avogadro *et al.* 2024] Roberto Avogadro, Fabio D'Adda, and Marco Cremaschi. Feature/vector entity retrieval and disambiguation techniques to create a supervised and unsupervised semantic table interpretation approach. *Knowledge-Based Systems*, 304:112447, 2024.

[Babaalla *et al.* 2024] Zakaria Babaalla, Hamza Abdelmalek, Abdeslam Jakimi, and Mohamed Oualla. Extraction of uml class diagrams using deep learning: Comparative study and critical analysis. *Procedia Computer Science*, 236:452–459, 2024. International Symposium on Green Technologies and Applications (ISGTA'2023).

[Bakos *et al.* 2025] Bence Bakos, Gábor Hidy, Bálint Csanády, Csaba Huszty, and András Lukács. Estimating room acoustic descriptors from bag-of-vectors representation with transformers. *Engineering Applications of Artificial Intelligence*, 145:110183, 2025.

[Bao *et al.* 2025] Tong Bao, Heng Zhang, and Chengzhi Zhang. Enhancing abstractive summarization of scientific papers using structure information. *Expert Systems with Applications*, 261:125529, 2025.

[Barmavat *et al.* 2025] Balaji Barmavat, M. Dhanaraju, K. Sreerama Murthy, Hari Krishna Madthala, Satya Krupa Prakash Karey, and Rajesh Palthya. Multilabel vulnerability classification in decentralized blockchain–based reputation system. *Journal of Software: Evolution and Process*, 37(4), 2025. Cited by: 0.

[Barrak *et al.* 2022] Mohamed Barrak, Bram Adams, and Amal Zouaq. *Toward a traceable, explainable, and fairJD/Resume recommendation system*, 02 2022.

[Başer *et al.* 2023] Furkan Başer, Oğuz Koç, and Sevtap Kestel. Credit risk evaluation using clustering based fuzzy classification method. *Expert Systems with Applications*, 223:119882, 03 2023.

[Bella *et al.* 2023] Giampaolo Bella, Cristian Daniele, and Mario Raciti. The aila methodology for automated and intelligent likelihood assignment in risk assessment. *IEEE ACCESS*, 11:26170–26183, 2023.

[Benefo *et al.* 2022] Edmund Benefo, Shraddha Karanth, and Abani K. Pradhan. Applications of advanced data analytic techniques in food safety and risk assessment. *CURRENT OPINION IN FOOD SCIENCE*, 48, DEC 2022.

[Bird and Lotfi 2024] Jordan J. Bird and Ahmad Lotfi. Customer service chatbot enhancement with attention-based transfer learning. *Knowledge-Based Systems*, 301:112293, 2024.

[Biswas *et al.* 2024] Baidyanath Biswas, Arunabha Mukhopadhyay, Ajay Kumar, and Dursun Delen. A hybrid framework using explainable ai (xai) in cyber-risk management for defence and recovery against phishing attacks. *Decision Support Systems*, 177:114102, 2024.

[Bonisoli *et al.* 2025] Giovanni Bonisoli, David Vilares, Federica Rollo, and Laura Po. Document-level event extraction from italian crime news using minimal data. *Knowledge-Based Systems*, 317:113386, 2025.

[Boyapati and Aygun 2025] Mallika Boyapati and Ramazan Aygun. Balancergnn: Balancer graph neural networks for imbalanced datasets: A case study on fraud detection. *Neural Networks*, 182:106926, 2025.

[Brahma *et al.* 2021] Arin Brahma, David M. Goldberg, Nohel Zaman, and Mariano Aloiso. Automated mortgage origination delay detection from textual conversations. *Decision Support Systems*, 140:113433, 2021.

[Brown *et al.* 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and Dario Amodei. *Language Models are Few-Shot Learners*, 05 2020.

[Buscemi and Proverbio 2024] Alessio Buscemi and Daniele Proverbio. *ChatGPT vs Gemini vs LLaMA on Multilingual Sentiment Analysis*, 05 2024.

[Cadeddu *et al.* 2024] Andrea Cadeddu, Alessandro Chessa, Vincenzo De Leo, Gianni Fenu, Enrico Motta, Francesco Osborne, Diego Reforgiato Recupero, Angelo Salatino, and Luca Secchi. A comparative analysis of knowledge injection strategies for large language models in the scholarly domain. *Engineering Applications of Artificial Intelligence*, 133:108166, 2024.

[Cai *et al.* 2023] Chengcheng Cai, Limin Pan, Xinshuai Li, Senlin Luo, and Zhouting Wu. A risk identification model for ict supply chain based on network embedding and text encoding. *Expert Systems with Applications*, 228:120459, 2023.

[Cerchiello *et al.* 2022] Paola Cerchiello, Giancarlo Nicola, Samuel Roennqvist, and Peter Sarlin. Assessing banks' distress using news and regular financial data. *FRONTIERS IN ARTIFICIAL INTELLIGENCE*, 5, JUN 2 2022.

[Chagnon *et al.* 2024] Eric Chagnon, Ronald Pandolfi, Jeffrey Donatelli, and Daniela Ushizima. Benchmarking topic models on scientific articles using berteley. *Natural Language Processing Journal*, 6:100044, 2024.

[Che *et al.* 2024] Wanliu Che, Zhao Wang, Cuiqing Jiang, and Mohammad Zoynul Abedin. Predicting financial distress using multimodal data: An attentive and regularized deep learning method. *Information Processing Management*, 61(4):103703, 2024.

[Chen *et al.* 2020] Weijia Chen, Zhijun Lu, Lijue You, Lingling Zhou, Jie Xu, and Ken Chen. Artificial intelligence-based multimodal risk assessment model for surgical site infection (amrams): Development and validation study. *JMIR MEDICAL INFORMATICS*, 8(6), JUN 2020.

[Chen *et al.* 2022] Yen-Liang Chen, Chen-Hsin Hsiao, and Chia-Chi Wu. An ensemble model for link prediction based on graph embedding. *Decision Support Systems*, 157:113753, 2022.

[Chen *et al.* 2023a] Xiaoyuan Chen, Turki Aljrees, Muhammad Umer, Hanen Karamti, Saba Tahir, Nihal Abuzinadah, Khaled Alnowaiser, Ala Eshmawi, Abdullah Mohamed, and Imran Ashraf. A novel approach for explicit song lyrics detection using machine and deep ensemble learning models. *PeerJ Computer Science*, 9:e1469, 08 2023.

[Chen *et al.* 2023b] Ze Chen, Wanting Ji, Linlin Ding, and Baoyan Song. Fine-grained document-level financial event argument extraction approach. *Engineering Applications of Artificial Intelligence*, 121:105943, 2023.

[Chen *et al.* 2025a] Junyi Chen, Leyuan Liu, and Fan Zhou. Do not wait: Preemptive rumor detection with cooperative llms and accessible social context. *Information Processing Management*, 62(3):103995, 2025.

[Chen *et al.* 2025b] Songlin Chen, Xiaoliang Chen, Duoqian Miao, Hongyun Zhang, Xiaolin Qin, and Peng Lu. Ada-uda: A transferable transformer framework for rumor detection using adversarial domain alignment within unsupervised domain adaptation. *Expert Systems with Applications*, 261:125487, 2025.

[Chen *et al.* 2025c] Xinyu Chen, Peifeng Li, and Qiaoming Zhu. Improving cross-document event coreference resolution by discourse coherence and structure. *Information Processing Management*, 62(4):104085, 2025.

[Chen *et al.* 2025d] Xueqin Chen, Xiaoyu Huang, Qiang Gao, Li Huang, and Guisong Liu. Enhancing text-centric fake news detection via external knowledge distillation from llms. *Neural Networks*, 187:107377, 2025.

[Chen *et al.* 2025e] Zhiling Chen, Hanning Chen, Mohsen Imani, Ruimin Chen, and Farhad Imani. Vision language model for interpretable and fine-grained detection of safety compliance in diverse workplaces. *Expert Systems with Applications*, 265:125769, 2025.

[Cheng *et al.* 2025] Yinghan Cheng, Qi Zhang, Chongyang Shi, Liang Xiao, Shufeng Hao, and Liang Hu. Cosd: Collaborative stance detection with contrastive heterogeneous topic graph learning. *Knowledge-Based Systems*, 317:113399, 2025.

[Chi *et al.* 2019] Sun Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. *How to Fine-Tune BERT for Text Classification?*, 05 2019.

[Chiu *et al.* 2024] Ming-Chuan Chiu, Pei-Yi Tai, and Chih-Yuan Chu. Developing a smart green supplier risk assessment system integrating natural language processing and life cycle assessment based on ahp framework: An empirical study. *RESOURCES CONSERVATION AND RECYCLING*, 207, AUG 2024.

[Chotirat and Meesad 2021] Saranlita Chotirat and Phayung Meesad. *Natural Language Processing with "More Than Words – BERT"*, pages 108–116. 06 2021.

[Chuang *et al.* 2022] Yi-Hsuan Chuang, Ja-Hwung Su, Ding-Hong Han, Yi-Wen Liao, Yeong-Chyi Lee, Yu-Fan Cheng, Tzung-Pei Hong, Katherine Shu-Min Li, Hsin-You Ou, Yi Lu, and Chih-Chi Wang. Effective natural language processing and interpretable machine learning for structuring ct liver-tumor reports. *IEEE ACCESS*, 10:116273–116286, 2022.

[Ciatto *et al.* 2025] Giovanni Ciatto, Andrea Agiollo, Matteo Magnini, and Andrea Omicini. Large language models as oracles for instantiating ontologies with domain-specific knowledge. *Knowledge-Based Systems*, 310:112940, 2025.

[Colakovic and Karakatič 2023] Ivona Colakovic and Sašo Karakatič. Fairboost: Boosting supervised learning for learning on multiple sensitive features. *Knowledge-Based Systems*, 280:110999, 2023.

[Colombo *et al.* 2025] Andrea Colombo, Anna Bernasconi, and Stefano Ceri. An llm-assisted etl pipeline to build a high-quality knowledge graph of the italian legislation. *Information Processing Management*, 62(4):104082, 2025.

[Cook and Karakuş 2024] Alec Cook and Oktay Karakuş. Llm-commentator: Novel fine-tuning strategies of large language models for automatic commentary generation using football event data. *Knowledge-Based Systems*, 300:112219, 2024.

[Corradi *et al.* 2022] Marie P.F. Corradi, Alyanne M. de Haan, Bernard Staumont, Aldert H. Piersma, Liesbet Geris, Raymond H.H. Pieters, Cyrille A.M. Krul, and Marc A.T. Teunis. Natural language processing in toxicology: Delineating adverse outcome pathways and guiding the application of new approach methodologies. *Biomaterials and Biosystems*, 7:100061, 2022.

[Corradi *et al.* 2024] Marie Corradi, Thomas Luechtefeld, Alyanne M. de Haan, Raymond Pieters, Jonathan H. Freedman, Tamara Vanhaecke, Mathieu Vinken, and Marc Teunis. The application of natural language processing for the extraction of mechanistic information in toxicology. *FRONTIERS IN TOXICOLOGY*, 6, MAY 10 2024.

[Cremaschi *et al.* 2025] Marco Cremaschi, Davide Ditolve, Cesare Curcio, Anna Panzeri, Andrea Spoto, and Andrea Maurino. Decoding the mind: A rag-llm on icd-11 for decision support in psychology. *Expert Systems with Applications*, 279:127191, 2025.

[Dai *et al.* 2025] Genan Dai, Jiayu Liao, Sicheng Zhao, Xianghua Fu, Xiaojiang Peng, Hu Huang, and Bowen Zhang. Large language model enhanced logic tensor network for stance detection. *Neural Networks*, 183:106956, 2025.

[Daphne *et al.* 2025] S. Daphne, V. Mary Anita Rajam, P. Hemanth, and Sundarrajan Dinesh. An ensemble patient graph framework for predictive modelling from electronic health records and medical notes. *DIAGNOSTICS*, 15(6), MAR 18 2025.

[Darwiesh *et al.* 2024] Abdelaziz Darwiesh, A.H. El-Baz, and Mohamed Elhoseny. Intelligent risk management system for enhancing performance of stock market applications. *Expert Systems with Applications*, 249:123493, 2024.

[Daudert 2021] Tobias Daudert. Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, 230:107389, 2021.

[de Fitero-Dominguez *et al.* 2024] David de Fitero-Dominguez, Eva Garcia-Lopez, Antonio Garcia-Cabot, and Jose-Javier Martinez-Herraiz. Enhanced automated code vulnerability repair using large language models. *Engineering Applications of Artificial Intelligence*, 138:109291, 2024.

[Debón and Domenech 2024] Ana Debón and Josep Domenech. Digital footprint approach for the study of competitiveness in wineries. *Expert Systems with Applications*, 257:125049, 2024.

[Deng *et al.* 2025] Yunteng Deng, Jia Song, Zhongliang Yang, Yilin Long, Li Zeng, and Linna Zhou. Diachronic semantic encoding based on pre-trained language model for temporal knowledge graph reasoning. *Knowledge-Based Systems*, 318:113479, 2025.

[Devi and Vallinayagi 2020] Uma Devi and Dr Vallinayagi. Blended sentiment polarity (bsp) for product review analysis. pages 4456–4466, 05 2020.

[Devlin *et al.* 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 10 2018.

[Ding *et al.* 2024] Zeyuan Ding, Zhihao Yang, and Hongfei Lin. A plug-and-play adapter for consistency identification in task-oriented dialogue systems. *Information Processing Management*, 61(3):103637, 2024.

[Dobbs *et al.* 2023] Matthew F. Dobbs, Alessia Mcgowan, Alexandria Selloni, Zarina Bilgrami, Cansu Sarac, Matthew Cotter, Shaynna N. Herrera, Guillermo A. Cecchi, Marianne Goodman, Cheryl M. Corcoran, and Agrima Srivastava. Linguistic correlates of suicidal ideation in youth at clinical high-risk for psychosis. *SCHIZOPHRENIA RESEARCH*, 259:20–27, SEP 2023.

[Dong and Chen 2024] Gaochen Dong and W. Chen. Blockwise compression of transformer-based models without retraining. *Neural Networks*, 171:423–428, 2024.

[Du and Tanaka-Ishii 2022] Xin Du and Kumiko Tanaka-Ishii. Stock portfolio selection balancing variance and tail risk via stock vector representation acquired from price data and texts. *Knowledge-Based Systems*, 249:108917, 2022.

[Du *et al.* 2023] Zhenjiao Du, Yixiang Xu, Changqi Liu, and Yonghui Li. plm4alg: Protein language model-based predictors for allergenic proteins and peptides. *JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY*, 72(1):752–760, DEC 19 2023.

[Du *et al.* 2024] Fei Du, Chunbing Bao, Qingchun Meng, and Yuqiao Hui. Text-based measure of esg risk exposure. *Procedia Computer Science*, 242:693–700, 2024. 11th International Conference on Information Technology and Quantitative Management (ITQM 2024).

[Du *et al.* 2025] Kunpeng Du, Xuan Zhang, Chen Gao, Weiyi Shang, Yubin Ma, Zhi Jin, and Linyu Li. Knowledge-enhanced prototypical network with graph structure and semantic information interaction for low-shot joint spoken language understanding. *Expert Systems with Applications*, 280:127569, 2025.

[Duan 2025] Mingli Duan. The signaling effect of tone: The influence of key audit matters' tone on bank lending decisions. *FINANCE RESEARCH LETTERS*, 76, APR 2025.

[Encord 2024] Encord. *From Vision to Edge: Meta's LLaMA 3.2 Explained*, September 30 2024. Accessed: 2024-11-19.

[Fan *et al.* 2024] Haolin Fan, Jerry Fuh, Wen Feng Lu, A. Senthil Kumar, and Bingbing Li. Unleashing the potential of large language models for knowledge augmentation: A practical experiment on incremental sheet forming. *Procedia Computer Science*, 232:1269–1278, 2024. 5th International Conference on Industry 4.0 and Smart Manufacturing (ISM 2023).

[Fan *et al.* 2025] Zhanling Fan, Chongcheng Chen, and Haifeng Luo. Time-uie: Tourism-oriented figure information model and unified information extraction via large language models. *Expert Systems with Applications*, 278:127434, 2025.

[Fatouros and Kouroumali 2023] Georgios Fatouros and Kalliopi Kouroumali. *Forex News Annotated Dataset for Sentiment Analysis*, May 2023.

[Fenerich *et al.* 2020] A. Fenerich, M. Steiner, Pedro Steiner Neto, E. Tochetto, D. Tsutsumi, Fernanda Assef, and B. Santos. Use of machine learning techniques in bank credit risk analysis. *Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería*, 36, 08 2020.

[Feng *et al.* 2024] Jianzhou Feng, Lazhi Zhao, Haonan Qin, Yiming Xu, and Ziqi Wang. Cadlra: A multi-charge prediction method based on the criminal act-driven law retrieval augmentation. *Engineering Applications of Artificial Intelligence*, 134:108619, 2024.

[Feng *et al.* 2025] Jianzhou Feng, Qin Wang, Huaxiao Qiu, and Lirong Liu. Retrieval in decoder benefits generative models for explainable complex question answering. *Neural Networks*, 181:106833, 2025.

[Ferrari and Spoletini 2025] Alessio Ferrari and Paola Spoletini. Formal requirements engineering and large language models: A two-way roadmap. *Information and Software Technology*, 181:107697, 2025.

[Fu *et al.* 2020] Xiangling Fu, Tianxiong Ouyang, Jinpeng Chen, and Xiaopeng Luo. Listening to the investors: A novel framework for online lending default prediction using deep learning neural networks. *Information Processing Management*, 57(4):102236, 2020.

[Gabriel *et al.* 2025] Rodney A. Gabriel, Brian H. Park, Chun-Nan Hsu, and Alvaro A. Macias. A review of leveraging artificial intelligence to predict persistent postoperative opioid use and opioid use disorder and its ethical considerations. *Current Pain and Headache Reports*, 29(1), 2025. Cited by: 0; All Open Access, Hybrid Gold Open Access.

[Galindo and Tamayo 2000] Jorge Galindo and Pablo Tamayo. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, 15:107–43, 02 2000.

[Gao *et al.* 2022] Guangxia Gao, Zhiwang Zhang, and Shiyong Kang. Multi-criteria linear optimization classifier with semantically weighted kernels for chinese word formation pattern prediction. *Procedia Computer Science*, 214:1506–1515, 2022. 9th International Conference on Information Technology and Quantitative Management.

[Gao *et al.* 2024a] Dehong Gao, Kaidi Chen, Ben Chen, Huangyu Dai, Linbo Jin, Wen Jiang, Wei Ning, Shanqing Yu, Qi Xuan, Xiaoyan Cai, Libin Yang, and Zhen Wang. Llms-based machine translation for e-commerce. *Expert Systems with Applications*, 258:125087, 2024.

[Gao *et al.* 2024b] Dehong Gao, Yufei Ma, Sen Liu, Mengfei Song, Linbo Jin, Wen Jiang, Xin Wang, Wei Ning, Shanqing Yu, Qi Xuan, Xiaoyan Cai, and Libin Yang. Fashiongpt: Llm instruction fine-tuning with multiple lora-adapter fusion. *Knowledge-Based Systems*, 299:112043, 2024.

[Gao *et al.* 2024c] Ling Gao, Hongda Zhang, Yiming Liu, Nan Sheng, Haotian Feng, and Hao Xu. Pgcl: Prompt guidance and self-supervised contrastive learning-based method for visual question answering. *Expert Systems with Applications*, 251:124011, 2024.

[García-Méndez *et al.* 2023] Silvia García-Méndez, Francisco de Arriba-Pérez, Ana Barros-Vila, and Francisco J. González-Castaño. Targeted aspect-based emotion analysis to detect opportunities and precaution in financial twitter messages. *Expert Systems with Applications*, 218:119611, 2023.

[Garg *et al.* 2025] Piyush Kumar Garg, Roshni Chakraborty, and Sourav Kumar Dandapat. Atsumm: Auxiliary information enhanced approach for abstractive disaster tweet summarization with sparse training data. *Knowledge-Based Systems*, 311:112969, 2025.

[Gatla 2023] Teja Gatla. Machine learning in credit risk assessment: Analyzing how machine learning models are transforming the assessment of credit risk for loans and credit cards. *Journal of Emerging Technologies and Innovative Research*, 10:k746–k750, 06 2023.

[Gou *et al.* 2025] Zhinan Gou, Yuchen Long, Jieli Sun, and Kai Gao. Tg-erc: Utilizing three generation models to handle emotion recognition in conversation tasks. *Expert Systems with Applications*, 268:126269, 2025.

[Guo *et al.* 2022] Yuchen Guo, Yao Li, and Yilei Qian. Local government debt risk assessment: A deep learning-based perspective. *INFORMATION PROCESSING & MANAGEMENT*, 59(3), MAY 2022.

[Guo *et al.* 2025] Hanghui Guo, Shimin Di, Zhangze Chen, Changfan Pan, Chaojun Meng, and Jia Zhu. Fgrcat: A fine-grained reasoning framework through causality and adversarial training. *Expert Systems with Applications*, 272:126688, 2025.

[Haghgou *et al.* 2024] Faezeh Haghgou, Ata Allah Taleizadeh, Mohsen Sadegh Amalnik, and Mostafa Hajiaghaei-Keshteli. Optimal pricing and patient satisfaction optimization for healthcare providers. *Expert Systems with Applications*, 247:123229, 2024.

[Hassija *et al.* 2020] Vikas Hassija, Gaurang Bansal, Vinay Chamola, Neeraj Kumar, and Mohsen Guizani. Secure lending: Blockchain and prospect theory-based decentralized credit scoring model. *IEEE Transactions on Network Science and Engineering*, PP, 03 2020.

[He *et al.* 2021] Tiancheng He, Joy Nolte Fong, Linda W. Moore, Chika F. Ezeana, David Victor, Mukul Divatia, Matthew Vasquez, R. Mark Ghobrial, and Stephen T. C. Wong. An imageomics and multi-network based deep learning model for risk assessment of liver transplantation for hepatocellular cancer. *COMPUTERIZED MEDICAL IMAGING AND GRAPHICS*, 89, APR 2021.

[He *et al.* 2024] Lu He, Samaneh Omranian, Susan McRoy, and Kai Zheng. *Using Large Language Models for sentiment analysis of health-related social media data: empirical evaluation and practical tips*, 03 2024.

[Heng and Subramanian 2022] Yi Heng and Preethi Subramanian. *A Systematic Review of Machine Learning and Explainable Artificial Intelligence (XAI) in Credit Risk Modelling*, pages 596–614. 10 2022.

[Hogenboom *et al.* 2016] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, and Emiel Caron. A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85:12–22, 2016.

[Hou *et al.* 2020] Jiaqi Hou, Xin Li, Rongchen Zhu, Chongqiang Zhu, Zeyu Wei, and Chao Zhang. A neural relation extraction model for distant supervision in counterterrorism scenario. *IEEE ACCESS*, 8:225088–225096, 2020.

[Hou *et al.* 2025a] Wenlong Hou, Weidong Zhao, Ning Jia, and Xianhui Liu. Low-resource knowledge graph completion based on knowledge distillation driven by large language models. *Applied Soft Computing*, 169:112622, 2025.

[Hou *et al.* 2025b] Zhaoguo Hou, Huawei Wang, Yubin Yue, Minglan Xiong, and Changchang Che. A novel method for cause portrait of aviation unsafe events based on hierarchical multi-task convolutional neural network. *Expert Systems with Applications*, 270:126466, 2025.

[Hu *et al.* 2024] Yujie Hu, Jianjun Yu, Kongmin Wang, and Fangyu Liao. Entity extraction for research and development project knowledge graph. *Procedia Computer Science*, 242:17–24, 2024. 11th International Conference on Information Technology and Quantitative Management (ITQM 2024).

[Hu *et al.* 2025] Peng Hu, Lei Sun, Cuiyun Hu, Xiuqing Mao, Song Guo, Jingwen Wang, and Miao Yu. Efficient and performant transformer private inference with heterogeneous attention mechanisms. *Applied Soft Computing*, 176:113150, 2025.

[Huang and Lei 2023] Da Huang and Fangyuan Lei. Temporal group-aware graph diffusion networks for dynamic link prediction. *Information Processing Management*, 60(3):103292, 2023.

[Huang *et al.* 2024] He Huang, Peixiang Cao, Yao Liu, Yan Lv, and Min Tong. Research on the application of large language model in electric power enterprise audit. *Procedia Computer Science*, 247:1388–1394, 2024. The 11th International Conference on Applications and Techniques in Cyber Intelligence.

[Huang *et al.* 2025a] Guanhua Huang, Bofei Gao, Jiaze Chen, Yuchen Zhang, and Zhouwang Yang. Enhancing fine-grained geographic named entity recognition by multi-scale siamese reconstruction network. *Engineering Applications of Artificial Intelligence*, 148:110287, 2025.

[Huang *et al.* 2025b] Shengzhi Huang, Yong Huang, Yinpeng Liu, Zhuoran Luo, and Wei Lu. Are large language models qualified reviewers in originality evaluation? *Information Processing Management*, 62(3):103973, 2025.

[Huang *et al.* 2025c] Shubin Huang, Yi Cai, Li Yuan, and Jiexin Wang. A knowledge-enhanced network for joint multimodal entity-relation extraction. *Information Processing Management*, 62(3):104033, 2025.

[Huang *et al.* 2025d] Yanlong Huang, Wenxin Tai, Fan Zhou, Qiang Gao, and Ting Zhong. Extracting key insights from earnings call transcript via information-theoretic contrastive learning. *Information Processing Management*, 62(3):103998, 2025.

[Huang *et al.* 2025e] Zhengze Huang, Yuqi Bai, Hengyu Liu, and Yun Lin. Multiscale feature modeling and interpretability analysis of the shap method for predicting the lifespan of landslide dams. *APPLIED SCIENCES-BASEL*, 15(5), MAR 2025.

[Hugging Face 2024] Hugging Face. *LLaMA 3.2 Model (1B) Release*, September 25 2024. Accessed: 2024-11-19.

[Hung *et al.* 2023] Ling-Chien Hung, Ying-Ying Su, Jui-Ming Sun, Wan-Ting Huang, and Sheng-Feng Sung. Clinical narratives as a predictor for prognosticating functional outcomes after intracerebral hemorrhage. *JOURNAL OF THE NEUROLOGICAL SCIENCES*, 453, OCT 15 2023.

[Hussain *et al.* 2024] Sadam Hussain, Mansoor Ali, Usman Naseem, Fahimeh Nezhadmoghadam, Munsif Ali Jatoi, T. Aaron Gulliver, and Jose Gerardo Tamez-Pena. Breast cancer risk prediction using machine learning: a systematic review. *FRONTIERS IN ONCOLOGY*, 14, MAR 20 2024.

[Hutto and Gilbert 2014] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8:216–225, 05 2014.

[Håkansson and Phillips-Wren 2024] Anne Håkansson and Gloria Phillips-Wren. Generative ai and large language models - benefits, drawbacks, future and recommendations. *Procedia Computer Science*, 246:5458–5468, 2024. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[Irarrázaval *et al.* 2021] María Elisa Irarrázaval, Sebastián Maldonado, Juan Pérez, and Carla Vairetti. Telecom traffic pumping analytics via explainable data science. *Decision Support Systems*, 150:113559, 2021. Interpretable Data Science For Decision Making.

[Islam and Zhang 2024] Md Saiful Islam and Long Zhang. *A Review on BERT: Language Understanding for Different Types of NLP Task*, 01 2024.

[Jeong and Gu 2024] Da Woon Jeong and Yeong Hyeon Gu. Pro trader rl: Reinforcement learning framework for generating trading knowledge by mimicking the decision-making patterns of professional traders. *Expert Systems with Applications*, 254:124465, 2024.

[Jeyakarthic and Ramesh 2023] M. Jeyakarthic and Ranjitha Ramesh. Genetic programming with dynamic bayesian network based credit risk assessment model. pages 845–850, 02 2023.

[Jiang *et al.* 2025] Jingchi Jiang, Lian Yan, Haifeng Liu, Zhenbo Xia, Haotian Wang, Yang Yang, and Yi Guan. Knowledge assimilation: Implementing knowledge-guided agricultural large language model. *Knowledge-Based Systems*, 314:113197, 2025.

[Jidkov *et al.* 2020] Vladislav Jidkov, Rami Abielmona, Alexander Teske, and Emil Petriu. Enabling maritime risk assessment using natural language processing-based deep learning techniques. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2469–2476, 2020.

[Jin and Zhang 2024] Jing Jin and Yongqing Zhang. Innovation in financial enterprise risk prediction model: A hybrid deep learning technique based on cnn-transformer-wt. *JOURNAL OF ORGANIZATIONAL AND END USER COMPUTING*, 36(1), 2024.

[Jin *et al.* 2025] Weiqiang Jin, Ningwei Wang, Tao Tao, Mengying Jiang, Yebei Xing, Biao Zhao, Hao Wu, Haibin Duan, and Guang Yang. A prompting multi-task learning-based veracity dissemination consistency reasoning augmentation for few-shot fake news detection. *Engineering Applications of Artificial Intelligence*, 144:110122, 2025.

[Jing *et al.* 2022] Sifeng Jing, Xiwei Liu, Xiaoyan Gong, Ying Tang, Gang Xiong, Sheng Liu, Shuguang Xiang, and Rongshan Bi. Correlation analysis and text classification of chemical accident cases based on word embedding. *PROCESS SAFETY AND ENVIRONMENTAL PROTECTION*, 158:698–710, FEB 2022.

[Jose *et al.* 2024] Sagar Jose, Khanh T.P Nguyen, Kamal Medjaher, Ryad Zemouri, Mélanie Lévesque, and Antoine Tahan. Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models. *Expert Systems with Applications*, 255:124603, 2024.

[Joshi *et al.* 2024] Ratnesh Kumar Joshi, Arindam Chatterjee, and Asif Ekbal. Saliency infused dialogue response generation: Improving task oriented text generation using feature attribution. *Expert Systems with Applications*, 255:124283, 2024.

[Jukna 2022] Daniels Jukna. Credit scoring methods comparison - a review. 05 2022.

[Kabal *et al.* 2024] Othmane Kabal, Mounira Harzallah, Fabrice Guillet, and Ryutaro Ichise. Enhancing domain-independent knowledge graph construction through openie cleaning and llms validation. *Procedia Computer Science*, 246:2617–2626, 2024. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[Kaefer *et al.* 2015] Florian Kaefer, Juliet Roper, and Paresha Sinha. A software-assisted qualitative content analysis of news articles: Example and reflections. *Forum Qualitative Sozialforschung*, 16, 05 2015.

[Kartal *et al.* 2024] Yusuf Kartal, E. Kaan Akdeniz, and Kemal Özkan. Automating modern code review processes with code similarity measurement. *Information and Software Technology*, 173:107490, 2024.

[Kim and Kim 2014] Joocheol Kim and Eunhwan Kim. An empirical analysis on credit risk models and its application. *Journal of Economic Financial Studies*, 2:14, 08 2014.

[Kim *et al.* 2024] Seulgi Kim, Seokwon Jeong, and Harksoo Kim. Bridge to better understanding: Syntax extension with virtual linking-phrase for natural language inference. *Knowledge-Based Systems*, 305:112608, 2024.

[Kirtac and Germano 2024] Kemal Kirtac and Guido Germano. Sentiment trading with large language models. *Finance Research Letters*, 62:105227, 2024.

[Korangi *et al.* 2023] Kamesh Korangi, Christophe Mues, and Cristian Bravo. A transformer-based model for default prediction in mid-cap corporate markets. *EUROPEAN JOURNAL OF OPERATIONAL RESEARCH*, 308(1):306–320, JUL 1 2023.

[Koroteev 2021] Mikhail Koroteev. *BERT: A Review of Applications in Natural Language Processing and Understanding*, 03 2021.

[Kumar *et al.* 2024] Harish Kumar, Anuradha Taluja, and Parsanjeet Kumar. A comprehensive analysis of lstm techniques for predicting financial market. *INTERNATIONAL JOURNAL OF FINANCIAL ENGINEERING*, 11(04), DEC 2024.

[Kumari 2023] Khushboo Kumari. *RoBERTa: A Modified BERT Model for NLP*, September 22 2023.

[Kuppan *et al.* 2024] Karthigeyan Kuppan, Deepak Bhaskar Acharya, and B. Divya. Foundational ai in insurance and real estate: A survey of applications, challenges, and future directions. *IEEE ACCESS*, 12:181282–181302, 2024.

[Lazaroiu *et al.* 2024] George Lazaroiu, Tom Gedeon, Elzbieta Rogalska, Katarina Valaskova, Marek Nagy, Hussam Musa, Katarina Zvarikova, Milos Poliak, Jakub Horak, Raluca Ionela Cretoiu, Tomas Krulicky, Luminita Ionescu, Catalin Popa, Lacramioara Rodica Hurloiu, Filip Nistor, Laurentia Georgeta Avram, and Viorica Braga. Digital twin-based cyber-physical manufacturing systems, extended reality metaverse enterprise and production management algorithms, and internet of things financial and labor market technologies in generative artificial intelligence economics. *OECONOMIA COPERNICANA*, 15(3):837–870, SEP 2024.

[Le *et al.* 2024] Duy Le, James Montgomery, Kenneth Kirkby, and Joel Scanlan. Using natural language processing to predict risk in electronic health records. *Studies in health technology and informatics*, 310:574–578, 01 2024.

[Lee and Lee 2024] Wonjong Lee and Seulki Lee. Development of a knowledge base for construction risk assessments using bert and graph models. *BUILDINGS*, 14(11), NOV 2024.

[Lessmann *et al.* 2015] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, (doi:10.1016/j.ejor.2015.05.030), 05 2015.

[Levkovich and Elyoseph 2023] Inbar Levkovich and Zohar Elyoseph. Suicide risk assessments through the eyes of chatgpt-3.5 versus chatgpt-4: Vignette study. *JMIR MENTAL HEALTH*, 10, 2023.

[Li and Dai 2024] Tianyu Li and Xiangyu Dai. Financial risk prediction and management using machine learning and natural language processing. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 15(6), JUN 2024.

[Li and He 2019] Feng Li and Zhuojing He. Credit risk clustering in a business group: Which matters more, systematic or idiosyncratic risk? *Cogent Economics Finance*, 7, 06 2019.

[Li *et al.* 2021] Shixuan Li, Wenxuan Shi, Jiancheng Wang, and Heshen Zhou. A deep learning-based approach to constructing a domain sentiment lexicon: a case study

in financial distress prediction. *INFORMATION PROCESSING & MANAGEMENT*, 58(5), SEP 2021.

[Li *et al.* 2022] Aihua Li, Qinyan Wei, Che Han, and Xinzhu Xing. Research on the construction of smart care question answering system based on knowledge graph. *Procedia Computer Science*, 214:1595–1602, 2022. 9th International Conference on Information Technology and Quantitative Management.

[Li *et al.* 2025a] Bo Li, Shuang Fan, Shaolin Zhu, and Lijie Wen. Cole: A collaborative legal expert prompting framework for large language models in law. *Knowledge-Based Systems*, 311:113052, 2025.

[Li *et al.* 2025b] Dong Li, Jintao Tang, Pancheng Wang, Shasha Li, and Ting Wang. Maximizing discrimination masking for faithful question answering with machine reading. *Information Processing Management*, 62(1):103915, 2025.

[Li *et al.* 2025c] Hui Li, Huanrong Liang, Yaohua Hu, and Xiaojie Liu. Technology convergence prediction based on temporal heterogeneous graph neural networks. *Information Processing Management*, 62(3):104034, 2025.

[Li *et al.* 2025d] Nan Li, Bo Kang, and Tijl De Bie. Llm4jobs: Unsupervised occupation extraction and standardization leveraging large language models. *Knowledge-Based Systems*, 316:113302, 2025.

[Li *et al.* 2025e] Zelin Li, Kehai Chen, Lemao Liu, Xuefeng Bai, Mingming Yang, Yang Xiang, and Min Zhang. Tf-attack: Transferable and fast adversarial attacks on large language models. *Knowledge-Based Systems*, 312:113117, 2025.

[Li *et al.* 2025f] Zihao Li, Lixin Zou, Chao Ma, and Chenliang Li. Efficient and explainable sequential recommendation with language model. *Information Processing Management*, 62(4):104122, 2025.

[Liang *et al.* 2022] Decui Liang, Bochun Yi, Wen Cao, and Qiang Zheng. Exploring ensemble oversampling method for imbalanced keyword extraction learning in policy text based on three-way decisions and smote. *Expert Systems with Applications*, 188:116051, 2022.

[Liang *et al.* 2024a] Du Liang, Zhang Rui, and Fu Yuguang. A robust evaluating strategy of tunnel deterioration using ensemble machine learning algorithms. *Engineering Applications of Artificial Intelligence*, 133:108364, 2024.

[Liang *et al.* 2024b] Xiao Liang, Di Wang, Haodi Zhong, Quan Wang, Ronghan Li, Rui Jia, and Bo Wan. Candidate-heuristic in-context learning: A new framework for enhancing medical visual question answering with llms. *Information Processing Management*, 61(5):103805, 2024.

[Lin and Liao 2024] Wei Lin and Li-Chuan Liao. Lexicon-based prompt for financial dimensional sentiment analysis. *Expert Systems with Applications*, 244:122936, 2024.

[Lin *et al.* 2020a] Wei-Chun Lin, Jimmy S. Chen, Michael F. Chiang, and Michelle R. Hribar. Applications of artificial intelligence to electronic health record data in

ophthalmology. *TRANSLATIONAL VISION SCIENCE & TECHNOLOGY*, 9(2), JAN 2020.

[Lin *et al.* 2020b] Yi Lin, Linjie Deng, Zhengmao Chen, Xiping Wu, Jianwei Zhang, and Bo Yang. A real-time atc safety monitoring framework using a deep learning approach. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, 21(11):4572–4581, NOV 2020.

[Lin *et al.* 2024a] Diefan Lin, Yi Wen, Weishi Wang, and Yan Su. Enhanced sentiment intensity regression through lora fine-tuning on llama 3. *IEEE Access*, PP:1–1, 01 2024.

[Lin *et al.* 2024b] Nankai Lin, Meiyu Zeng, Xingming Liao, Weizhong Liu, Aimin Yang, and Dong Zhou. Addressing class-imbalance challenges in cross-lingual aspect-based sentiment analysis: Dynamic weighted loss and anti-decoupling. *Expert Systems with Applications*, 257:125059, 2024.

[Lin *et al.* 2025] Feng Lin, Hanling Yi, Yifan Yang, Hongbin Li, Xiaotian Yu, Guangming Lu, and Rong Xiao. Bita: Bi-directional tuning for lossless acceleration in large language models. *Expert Systems with Applications*, 279:127305, 2025.

[Ling *et al.* 2025] Yuchen Ling, Shengcheng Yu, Chunrong Fang, Guobin Pan, Jun Wang, and Jia Liu. Redefining crowdsourced test report prioritization: An innovative approach with large language model. *Information and Software Technology*, 179:107629, 2025.

[Lis *et al.* 2023] Szymon Lis, Mariusz Kubkowski, Olimpia Borkowska, Dobromił Serwa, and Jarosław Kurpanik. *Analyzing Credit Risk Model Problems through NLP-Based Clustering and Machine Learning: Insights from Validation Reports*, 06 2023.

[Liu and Yang 2022] Chang Liu and Shiwu Yang. Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *EXPERT SYSTEMS WITH APPLICATIONS*, 207, NOV 30 2022.

[Liu *et al.* 2019a] Weidong Liu, Caixia Qin, Kun Gao, L. Heng, Zuen Qin, Yafei Cao, and Wen Si. Research on medical data feature extraction and intelligent recognition technology based on convolutional neural network. *IEEE ACCESS*, 7:150157–150167, 2019.

[Liu *et al.* 2019b] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[Liu *et al.* 2021] Ximeng Liu, Lehui Xie, Yaopeng Wang, Jian Zou, Jinbo Xiong, Zuobin Ying, and Athanasios V. Vasilakos. Privacy and security issues in deep learning: A survey. *IEEE ACCESS*, 9:4566–4593, 2021.

[Liu *et al.* 2023a] Keyang Liu, Baoping Cai, Qibing Wu, Mingxin Chen, Chao Yang, Javed Akbar Khan, Chenyushu Wang, Hasini Vidumini Weerawarna Pattiyakumbura, Weifeng Ge, and Yonghong Liu. Risk identification and assessment methods

of offshore platform equipment and operations. *PROCESS SAFETY AND ENVIRONMENTAL PROTECTION*, 177:1415–1430, SEP 2023.

[Liu *et al.* 2023b] Tianyang Liu, Chong Wang, Kun Huang, Peng Liang, Beiqi Zhang, Maya Daneva, and Marten van Sinderen. Rosematcher: Identifying the impact of user reviews on app updates. *Information and Software Technology*, 161:107261, 2023.

[Liu *et al.* 2024a] Jingping Liu, Xianyang Tian, Hanwen Tong, Chenhao Xie, Tong Ruan, Lin Cong, Baohua Wu, and Haofen Wang. Enhancing chinese abbreviation prediction with llm generation and contrastive evaluation. *Information Processing Management*, 61(4):103768, 2024.

[Liu *et al.* 2024b] Liu Liu, Ming Liu, Shanshan Liu, and Kun Ding. Event extraction as machine reading comprehension with question-context bridging. *Knowledge-Based Systems*, 299:112041, 2024.

[Liu *et al.* 2024c] Ruiheng Liu, Yu Zhang, Bailong Yang, Qi Shi, and Luogeng Tian. Robust and resource-efficient table-based fact verification through multi-aspect adversarial contrastive learning. *Information Processing Management*, 61(6):103853, 2024.

[Liu *et al.* 2024d] Wei Liu, Zheng Huang, Chao Wang, Yan Peng, and Shaorong Xie. Eglr: Two-staged explanation generation and language reasoning framework for commonsense question answering. *Knowledge-Based Systems*, 286:111411, 2024.

[Liu *et al.* 2024e] Yu Liu, Duantengchuan Li, Kaili Wang, Zhuoran Xiong, Fobo Shi, Jian Wang, Bing Li, and Bo Hang. Are llms good at structured outputs? a benchmark for evaluating structured output capabilities in llms. *Information Processing Management*, 61(5):103809, 2024.

[Liu *et al.* 2024f] Zhongkun Liu, Zhumin Chen, Zhaochun Ren, Shen Gao, Jun Ma, and Pengjie Ren. Center-retained fine-tuning for conversational question ranking through unsupervised center identification. *Information Processing Management*, 61(2):103578, 2024.

[Liu *et al.* 2025a] Alfred Z. Liu, Angela Xia Liu, and Kexin Zhao. Disclosure of it-related risk factors in corporate filings. *Decision Support Systems*, 190:114403, 2025.

[Liu *et al.* 2025b] Caizheng Liu, Zhengyu Zhu, Wanming Hao, and Gangcan Sun. Heterogeneous multivariate time series imputation by transformer model with missing position encoding. *Expert Systems with Applications*, 271:126435, 2025.

[Liu *et al.* 2025c] Haijiang Liu, Yong Cao, Xun Wu, Chen Qiu, Jinguang Gu, Maofu Liu, and Daniel Hershcovich. Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation. *Information Processing Management*, 62(4):104099, 2025.

[Liu *et al.* 2025d] Haitao Liu, Xianwei Xin, Weiming Peng, Jihua Song, and Jingbo Sun. Concept-driven knowledge distillation and pseudo label generation for continual named entity recognition. *Expert Systems with Applications*, 270:126546, 2025.

[Liu *et al.* 2025e] Jiandong Liu, Jianfeng Ren, Ruibin Bai, Zibo Zhang, and Zheng Lu. A cascaded retrieval-while-reasoning multi-document comprehension framework with incremental attention for medical question answering. *Expert Systems with Applications*, 265:125701, 2025.

[Liu *et al.* 2025f] Qingli Liu, Fan Li, Kam K.H. Ng, Jiashu Han, and Shanshan Feng. Accident investigation via llms reasoning: Hfacs-guided chain-of-thoughts enhance general aviation safety. *Expert Systems with Applications*, 269:126422, 2025.

[Liu *et al.* 2025g] Ran Liu, Xian-Ling Mao, and Heyan Huang. Dscisum: Detailed summarization of long scientific documents. *Knowledge-Based Systems*, 317:113409, 2025.

[Liu *et al.* 2025h] Shiqi Liu, Weizheng Kong, Zhi Liu, Jianwen Sun, Sannyuya Liu, and Dragan Gašević. Dual-view cross attention enhanced semi-supervised learning method for discourse cognitive engagement classification in online course discussions. *Expert Systems with Applications*, 278:127339, 2025.

[Liu *et al.* 2025i] Siqi Liu, Zhusen Liu, Donglong Chen, Wangchen Dai, Lu Zhou, Zhe Liu, Ray C. C. Cheung, and Çetin Kaya Koç. Mlformer: a high performance mpc linear inference framework for transformers. *Journal of Cryptographic Engineering*, 15(1), 2025. Cited by: 0.

[Liu *et al.* 2025j] Yinhao Liu, Zhenyu Kuang, Hongyang Zhang, Chen Li, Feifei Li, and Xinghao Ding. Prada: Prompt-guided representation alignment and dynamic adaption for time series forecasting. *Knowledge-Based Systems*, page 113478, 2025.

[Luna-Jiménez *et al.* 2024] Cristina Luna-Jiménez, Manuel Gil-Martín, Luis Fernando D'Haro, Fernando Fernández-Martínez, and Rubén San-Segundo. Evaluating emotional and subjective responses in synthetic art-related dialogues: A multistage framework with large language models. *Expert Systems with Applications*, 255:124524, 2024.

[López Espejel *et al.* 2023] Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032, 2023.

[Mahbobi *et al.* 2021] Mo Mahbobi, Salman Kimiagari, and Marriappan Vasudevan Marriappan Vasudevan. Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*, 330:1–29, 07 2021.

[Mahgoub *et al.* 2024] Abeer Mahgoub, Ghada Khoriba, and ElHassan Anas ElSabry. Mathematical problem solving in arabic: Assessing large language models. *Procedia Computer Science*, 244:86–95, 2024. 6th International Conference on AI in Computational Linguistics.

[Mai *et al.* 2024] Chengcheng Mai, Yuxiang Wang, Ziyu Gong, Hanxiang Wang, Kaiwen Luo, Chunfeng Yuan, and Yihua Huang. iterprompt: An iterative prompt-tuning method for nested relation extraction with dynamic assignment strategy. *Expert Systems with Applications*, 254:124335, 2024.

[Maldonado *et al.* 2019] Sebastián Maldonado, Julio López, and Carla Vairetti. An alternative smote oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, 76:380–389, 2019.

[Malhotra and Jindal 2024] Anshu Malhotra and Rajni Jindal. Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks. *COGNITIVE SYSTEMS RESEARCH*, 84, MAR 2024.

[Malik *et al.* 2024] Pankaj Malik, Anoushka Anand, Anmol Baliyan, Anant Dongre, and Palak Panwar. Credit risk assessment and fraud detection in financial transactions using machine learning. Volume 10:433–440, 04 2024.

[Manjotho *et al.* 2025] Ali Asghar Manjotho, Tekie Tsegay Tewolde, Ramadhani Ally Duma, and Zhendong Niu. Llm-guided fuzzy kinematic modeling for resolving kinematic uncertainties and linguistic ambiguities in text-to-motion generation. *Expert Systems with Applications*, 279:127283, 2025.

[Mann *et al.* 2023] Suman Mann, Jyoti Arora, Mudita Bhatia, Ritika Sharma, and Rewangi Taragi. *Twitter Sentiment Analysis Using Enhanced BERT*, pages 263–271. 01 2023.

[Marulli *et al.* 2024] Fiammetta Marulli, Lelio Campanile, Maria Stella de Biase, Stefano Marrone, Laura Verde, and Marianna Bifulco. Understanding readability of large language models output: An empirical analysis. *Procedia Computer Science*, 246:5273–5282, 2024. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[Masud *et al.* 2025] Mohammed Tanvir Masud, Nickolaos Koroniotis, Marwa Keshk, Benjamin Turnbull, Shabnam Kasra Kermanshahi, and Nour Moustafa. Generative fuzzer-driven vulnerability detection in the internet of things networks. *APPLIED SOFT COMPUTING*, 174, APR 2025.

[Mei *et al.* 2024] Yingxue Mei, Zicai Jin, Weiguo Ma, Yingjun Ma, Ning Deng, Zhiyuan Fan, and Shujun Wei. Optimizing acute coronary syndrome patient treatment: Leveraging gated transformer models for precise risk prediction and management. *BIOENGINEERING-BASEL*, 11(6), JUN 2024.

[Meng *et al.* 2022] Zaiqiao Meng, Anya Okhmatovskaia, Maxime Polleri, Yannan Shen, Guido Powell, Zihao Fu, Iris Ganser, Meiru Zhang, Nicholas B. King, David Buckeridge, and Nigel Collier. Biocaster in 2021: automatic disease outbreaks detection from global news media. *BIOINFORMATICS*, 38(18):4446–4448, SEP 15 2022.

[Mersha *et al.* 2025] Melkamu Abay Mersha, Mesay Gemeda Yigezu, and Jugal Kalita. Evaluating the effectiveness of xai techniques for encoder-based language models. *Knowledge-Based Systems*, 310:113042, 2025.

[Merton 2019] Robert Merton. *On the Pricing of Corporate Debt: The Risk Structure of Interest Rates*, pages 79–102. 03 2019.

[Miao *et al.* 2025] Shumei Miao, Pei Ji, Yongqian Zhu, Haoyu Meng, Mang Jing, Rongrong Sheng, Xiaoliang Zhang, Hailong Ding, Jianjun Guo, Wen Gao, Guanyu Yang, and Yun Liu. The construction and application of a clinical decision support system for cardiovascular diseases: Multimodal data-driven development and validation study. *JMIR MEDICAL INFORMATICS*, 13, 2025.

[Moscato *et al.* 2024] Vincenzo Moscato, Marco Postiglione, Giancarlo Sperlì, and Andrea Vignali. Aldaner: Active learning based data augmentation for named entity recognition. *Knowledge-Based Systems*, 305:112682, 2024.

[Mu *et al.* 2024] Yida Mu, Pu Niu, Kalina Bontcheva, and Nikolaos Aletras. Predicting and analyzing the popularity of false rumors in weibo. *Expert Systems with Applications*, 243:122791, 2024.

[Nafa *et al.* 2024] Youcef Nafa, Qun Chen, Boyi Hou, and Zhanhuai Li. Adaptive deep learning for entity disambiguation via knowledge-based risk analysis. *Expert Systems with Applications*, 238:122342, 2024.

[Nair *et al.* 2024] Michael Nair, Laila Abd-Elmegid, and Mohamed Marie. Sentiment analysis model for cryptocurrency tweets using different deep learning techniques. *Journal of Intelligent Systems*, 33, 03 2024.

[Najem *et al.* 2022] Rihab Najem, Meryem Fakhouri Amr, Ayoub Bahnasse, and Mohamed Talea. Artificial intelligence for digital finance, axes and techniques. *Procedia Computer Science*, 203:633–638, 2022. 17th International Conference on Future Networks and Communications / 19th International Conference on Mobile Systems and Pervasive Computing / 12th International Conference on Sustainable Energy Information Technology (FNC/MobiSPC/SEIT 2022), August 9-11, 2022, Niagara Falls, Ontario, Canada.

[Nguyen *et al.* 2023] Dieu-Hien Nguyen, Nguyen-Khang Le, and Le-Minh Nguyen. Viwiqa: Efficient end-to-end vietnamese wikipedia-based open-domain question-answering systems for single-hop and multi-hop questions. *Information Processing Management*, 60(6):103514, 2023.

[Nguyen *et al.* 2025a] Chau Nguyen, Phuong Nguyen, and Le-Minh Nguyen. Retrieve–revise–refine: A novel framework for retrieval of concise entailing legal article set. *Information Processing Management*, 62(1):103949, 2025.

[Nguyen *et al.* 2025b] Xuan-Bang Nguyen, Xuan-Hieu Phan, and Massimo Piccardi. Fine-tuning text-to-sql models with reinforcement-learning training objectives. *Natural Language Processing Journal*, 10:100135, 2025.

[Nimmy *et al.* 2022] Sonia Farhana Nimmy, Omar K. Hussain, Ripon K. Chakrabortty, Farookh Khadeer Hussain, and Morteza Saberi. Explainability in supply chain operational risk management: A systematic literature review. *Knowledge-Based Systems*, 235:107587, 2022.

[Noriega *et al.* 2023] Jomark Noriega, Luis Rivera, and Jose Herrera. *Machine Learning for Credit Risk Prediction: A Systematic Literature Review*, 08 2023.

[Obeid *et al.* 2020] Jihad S. Obeid, Matthew Davis, Matthew Turner, Stephane M. Meystre, Paul M. Heider, Edward C. O'Bryan, and Leslie A. Lenert. An artificial intelligence approach to covid-19 infection risk assessment in virtual visits: A case report. *JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*, 27(8):1321–1325, AUG 2020.

[O'Keefe 2024] John O'Keefe. The use of machine learning for credit risk assessment. 05 2024.

[OU and ZHANG 2025] Ziquan OU and Zijun ZHANG. Audan: An adversarial unsupervised domain adaptation network informed decision-making in audit risk assessment. *Knowledge-Based Systems*, 314:113171, 2025.

[Oualid *et al.* 2022] Adil Oualid, Abderrahim Hansali, Youssef Balouki, and Lahcen Moumoun. *Application of Machine Learning Techniques for Credit Risk Management: A Survey*, pages 180–191. 01 2022.

[Ouyang *et al.* 2021] Zi-sheng Ouyang, Xi-te Yang, and Yongzeng Lai. Systemic financial risk early warning of financial market in china using attention-lstm model. *NORTH AMERICAN JOURNAL OF ECONOMICS AND FINANCE*, 56, APR 2021.

[Pakdaman *et al.* 2025] Zahra Pakdaman, Abbas Koochari, and Arash Sharifi. Content-aware sentiment understanding: cross-modal analysis with encoder-decoder architectures. *Journal of Computational Social Science*, 8, 02 2025.

[Palenzuela Baena *et al.* 2019] Jose Antonio Palenzuela Baena, Silvia Scifoni, Maria Marsella, Gianfilippo De Astis, and Clemente Irigaray Fernandez. Landslide susceptibility mapping on the islands of vulcano and lipari (aeolian archipelago, italy), using a multi-classification approach on conditioning factors and a modified gis matrix method for areas lacking in a landslide inventory. *LANDSLIDES*, 16(5):969–982, MAY 2019.

[Pan *et al.* 2025] Yue Pan, Chen Lyu, Zhenyu Yang, Lantian Li, Qi Liu, and Xiuting Shao. E-code: Mastering efficient code generation through pretrained models and expert encoder group. *Information and Software Technology*, 178:107602, 2025.

[Pandey *et al.* 2017] Trilok Pandey, Alok Jagadev, Suman Mohapatra, and Satchidananda Dehuri. Credit risk analysis using machine learning classifiers. pages 1850–1854, 08 2017.

[Papatheodosiou *et al.* 2024] Persephone Papatheodosiou, Dimitrios P. Panagoulias, Maria Virvou, George A. Tsihrintzis, Anastasios Bonakis, and Dimitrios Dikeos. Leveraging artificial intelligence for personalised insomnia-sleep calibration via the big five personality traits. *Procedia Computer Science*, 246:2539–2548, 2024. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[Paraskevopoulos *et al.* 2022] Sotirios Paraskevopoulos, Patrick Smeets, Xin Tian, and Gertjan Medema. Using artificial intelligence to extract information on pathogen characteristics from scientific publications. *INTERNATIONAL JOURNAL OF HYGIENE AND ENVIRONMENTAL HEALTH*, 245, AUG 2022.

[Parsa *et al.* 2025] Shyon Parsa, Priyansh Shah, Ritu Doijad, and Fatima Rodriguez. Artificial intelligence in ischemic heart disease prevention. *Current Cardiology Reports*, 27(1), 2025. Cited by: 0.

[Patel 2023] Kaushikkumar Patel. Credit card analytics: A review of fraud detection and risk assessment techniques. *International Journal of Biotech Trends and Technology*, 71:69–79, 10 2023.

[Pathak *et al.* 2023] Kirtiman Pathak, Christophe Rougeaux, and Himanshu Singh. *Navigating Economic Uncertainty: New Guidance for Credit Risk Management*, April 2023.

[Patsakis *et al.* 2024] Constantinos Patsakis, Fran Casino, and Nikolaos Lykousas. Assessing llms in malicious code deobfuscation of real-world malware campaigns. *Expert Systems with Applications*, 256:124912, 2024.

[Pellegrino 2025] Roberta Pellegrino. Risk management of circular economy: a framework based on interpretive structural model. *Procedia Computer Science*, 253:683–690, 2025. 6th International Conference on Industry 4.0 and Smart Manufacturing.

[Pham and Han 2023] Hieu T. T. L. Pham and SangUk Han. Natural language processing with multitask classification for semantic prediction of risk-handling actions in construction contracts. *JOURNAL OF COMPUTING IN CIVIL ENGINEERING*, 37(6), NOV 1 2023.

[Phan *et al.* 2023] Minh Phan, Arno De Caigny, and Kristof Coussement. A decision support framework to incorporate textual data for early student dropout prediction in higher education. *Decision Support Systems*, 168:113940, 2023.

[Prabhakar and Lee 2022] Sunil Kumar Prabhakar and Seong-Whan Lee. Enic: Ensemble and nature inclined classification with sparse depiction based deep and transfer learning for biosignal classification. *Applied Soft Computing*, 117:108416, 2022.

[Płonka *et al.* 2025] Mateusz Płonka, Krzysztof Kocot, Kacper Hołda, Krzysztof Daniec, and Aleksander Nawrat. A comparative evaluation of the effectiveness of document splitters for large language models in legal contexts. *Expert Systems with Applications*, 272:126711, 2025.

[Qin *et al.* 2023] Wenke Qin, Lang Chen, Boyi Zhang, Weiya Chen, and Hanbin Luo. Neodescriber: An image-to-text model for automatic style description of neoclassical architecture. *Expert Systems with Applications*, 231:120706, 2023.

[R *et al.* 2023] Sudheesh R, Muhammad Mujahid, Furqan Rustam, Rahman Shafique, Venkata Chunduri, Mònica Gracia, Julién Ballester, Isabel De la Torre Díez, and

Imran Ashraf. Analyzing sentiments regarding chatgpt using novel bert model. *Information*, 14:474, 08 2023.

[Raeesi *et al.* 2020] Mohsen Raeesi, Masoud Asadpour, and Azadeh Shakery. Swash: A collective personal name matching framework. *Expert Systems with Applications*, 147:113115, 2020.

[Rahaman *et al.* 2024] Md Facklasur Rahaman, Mohtasin Golam, Md Raihan Subhan, Esmot Ara Tuli, Dong-Seong Kim, and Jae-Min Lee. Meta-governance: Blockchain-driven metaverse platform for mitigating misbehavior using smart contract and ai. *IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT*, 21(4):4024–4038, AUG 2024.

[Rehman *et al.* 2025] Mohammad Zia Ur Rehman, Sufyaan Zahoor, Areeb Manzoor, Musharaf Maqbool, and Nagendra Kumar. A context-aware attention and graph neural network-based multimodal framework for misogyny detection. *Information Processing Management*, 62(1):103895, 2025.

[Rodríguez-Ibánez *et al.* 2023] Margarita Rodríguez-Ibánez, Antonio Casanez-Ventura, Félix Castejón-Mateos, and Pedro-Manuel Cuenca-Jiménez. A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 223:119862, 03 2023.

[Rogojan *et al.* 2023] Luana Rogojan, Andreea Croicu, and Laura Iancu. Modern approaches in credit risk modeling: A literature review. *Proceedings of the International Conference on Business Excellence*, 17:1617–1627, 07 2023.

[Ronnqvist and Sarlin 2017] Samuel Ronnqvist and Peter Sarlin. Bank distress in the news: Describing events through deep learning. *NEUROCOMPUTING*, 264(SI):57–70, NOV 15 2017.

[Roshanzamir *et al.* 2021] Alireza Roshanzamir, Hamid Aghajan, and Mahdieh Soleymani Baghshah. Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech. *BMC MEDICAL INFORMATICS AND DECISION MAKING*, 21(1), MAR 9 2021.

[Roumeliotis *et al.* 2024] Konstantinos Roumeliotis, Nikolaos Tselikas, and Dimitrios Nasiopoulos. Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056, 03 2024.

[Rustam *et al.* 2024] Didi Rustam, Adang Suhendra, Suryadi Harmanto, Ruddy Suhatril, Dwi Saputra, Rusdan Tafsili, and Rizky Prasetya. Development of classification method for lecturer area of expertise based on scientific publication using bert. *International Journal on Advanced Science, Engineering and Information Technology*, 14:894–905, 06 2024.

[Safari and Ghaemi 2025] Ashkan Safari and Sehraneh Ghaemi. Neurofuzzyman: A hybrid neuro-fuzzy bilstm stacked ensemble model for financial forecasting and analysis: Dataset case studies on jpmorgan, amzn and tsla. *Expert Systems with Applications*, 266:126037, 2025.

[Safitra *et al.* 2024] Muhammad Fakhrul Safitra, Muharman Lubis, Tien Fabrianti Kusumasari, and Deyana Prastika Putri. Advancements in artificial intelligence and data science: Models, applications, and challenges. *Procedia Computer Science*, 234:381–388, 2024. Seventh Information Systems International Conference (ISICO 2023).

[Salihu and Shehu 2020] Armend Salihu and Visar Shehu. Data mining based classifiers for credit risk analysis. *Managing Global Transitions*, 18:147–167, 06 2020.

[Sanz-Guerrero and Arroyo 2024] Mario Sanz-Guerrero and Javier Arroyo. Credit risk meets large language models: Building a risk indicator from loan descriptions in p2p lending. *arXiv preprint arXiv:2401.16458*, 2024.

[Sanz-Guerrero and Arroyo 2025a] Mario Sanz-Guerrero and Javier Arroyo. Credit risk meets large language models: Building a risk indicator from loan descriptions in p2p lending. *Inteligencia Artificial*, 28(75):220 – 248, 2025. Cited by: 0; All Open Access, Gold Open Access.

[Sanz-Guerrero and Arroyo 2025b] Mario Sanz-Guerrero and Javier Arroyo. Credit risk meets large language models: Building a risk indicator from loan descriptions in p2p lending. *INTELIGENCIA ARTIFICIAL-IBEROAMERICAN JOURNAL OF ARTIFICIAL INTELLIGENCE*, 28(75):220–247, JUN 2025.

[Sarfati *et al.* 2024] Raphaël Sarfati, Toni Liu, Nicolas Boullé, and Christopher Earls. *Lines of Thought in Large Language Models*, 10 2024.

[Sarvazyan *et al.* 2024] Areg Mikael Sarvazyan, José Ángel González, and Marc Franco-Salvador. Textmachina: Seamless generation of machine-generated text datasets. *Procedia Computer Science*, 246:566–575, 2024. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

[Schuerkamp *et al.* 2025] Ryan Schuerkamp, Hannah Ahlstrom, and Philippe J. Giabbanelli. Automatically resolving conflicts between expert systems: An experimental approach using large language models and fuzzy cognitive maps from participatory modeling studies. *Knowledge-Based Systems*, 313:113151, 2025.

[Shafee *et al.* 2025] Samaneh Shafee, Alysson Bessani, and Pedro M. Ferreira. Evaluation of llm-based chatbots for osint-based cyber threat awareness. *Expert Systems with Applications*, 261:125509, 2025.

[Shang *et al.* 2023] Lili Shang, Hao Xi, Jiaojiao Hua, Huayun Tang, and Jilei Zhou. A lexicon enhanced collaborative network for targeted financial sentiment analysis. *Information Processing Management*, 60(2):103187, 2023.

[Shao *et al.* 2024] Minghao Shao, Abdul Basit, Ramesh Karri, and Muhammad Shafique. *Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges*, 09 2024.

[Sharma and Bhalla 2022] Kirti Sharma and Rajni Bhalla. Decision support machine-a hybrid model for sentiment analysis of news headlines of stock market. *Inter-*

national journal of electrical and computer engineering systems, 13:791–798, 12 2022.

[Shen and Wu 2025a] Chuanhe Shen and Junzhe Wu. Research on credit risk of listed companies: a hybrid model based on tcn and dilateformer. *SCIENTIFIC REPORTS*, 15(1), JAN 21 2025.

[Shen and Wu 2025b] Chuanhe Shen and Junzhe Wu. Research on credit risk of listed companies: a hybrid model based on tcn and dilateformer. *Scientific Reports*, 15(1), 2025. Cited by: 0; All Open Access, Gold Open Access.

[Shen *et al.* 2025] Qingyang Shen, Xiaozhi Zhang, Haomin Ren, Quan Guo, and Zhang Yi. Knowledge-embedded large language models for emergency triage. *Knowledge-Based Systems*, page 113431, 2025.

[Shi and Zheng 2021] Dejia Shi and Hanzhong Zheng. A mortality risk assessment approach on icu patients clinical medication events using deep learning. *CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES*, 128(1):161–181, 2021.

[Shi *et al.* 2022] Si Shi, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau. Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications*, 34, 07 2022.

[Shi *et al.* 2024] Donghui Shi, Zhigang Li, Jozef Zurada, Andrew Manikas, Jian Guan, and Pawel Weichbroth. Ontology-based text convolution neural network (textcnn) for prediction of construction accidents. *KNOWLEDGE AND INFORMATION SYSTEMS*, 66(4):2651–2681, APR 2024.

[Shrestha *et al.* 2025] Rabin Shrestha, Taewoo Ko, and Jeehee Lee. Quantifying project uncertainties: Leveraging historical bid and change order data for automated detection of cost and schedule impacts in new projects. *Journal of Construction Engineering and Management*, 151(4), 2025. Cited by: 0.

[Shung 2021] Dennis L. Shung. Advancing care for acute gastrointestinal bleeding using artificial intelligence. *JOURNAL OF GASTROENTEROLOGY AND HEPATOLOGY*, 36(2, SI):273–278, FEB 2021.

[Šmíd *et al.* 2024] Jakub Šmíd, Pavel Priban, and Pavel Kral. LLaMA-based models for aspect-based sentiment analysis. In Orphée De Clercq, Valentin Barriere, Jeremy Barnes, Roman Klinger, João Sedoc, and Shabnam Tafreshi, editors, *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 63–70, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Song *et al.* 2024a] Yunan Song, Xuewei An, Wenkai Zhang, Kaiyang Sun, and Anqi Zhang. Catalyzing financial risk control excellence: A novel fusion model - pso-boost-trans. *JOURNAL OF ORGANIZATIONAL AND END USER COMPUTING*, 36(1), 2024.

[Song *et al.* 2024b] Yunan Song, Huaqing Du, Tianyu Piao, and Hongyu Shi. Research on financial risk intelligent monitoring and early warning model based on lstm,

transformer, and deep learning. *JOURNAL OF ORGANIZATIONAL AND END USER COMPUTING*, 36(1), 2024.

[Song *et al.* 2025a] Junru Song, Tianlei Chen, Yang Yang, and Feifei Wang. Hetero-htc: Enhancing hierarchical text classification via heterogeneity encoding of label hierarchy. *Expert Systems with Applications*, 271:126558, 2025.

[Song *et al.* 2025b] Sangmin Song, Juhyoung Park, Juhwan Choi, Junho Lee, Kyohoon Jin, and YoungBin Kim. Korean football in-game conversation state tracking dataset for dialogue and turn level evaluation. *Engineering Applications of Artificial Intelligence*, 139:109572, 2025.

[Sridevi and Thambusamy 2024] Perumal Sridevi and Velmurugan Thambusamy. Twitter sentiment analysis of covid-19 vaccination integrating senticnet-7 and sentiwordnet-adjusted vader models. 16:162–183, 01 2024.

[Stefan and Brad 2024] Repede Stefan and Remus Brad. Llama 3 vs. state-of-the-art large language models: Performance in detecting nuanced fake news. *Computers*, 13:292, 11 2024.

[Stevenson *et al.* 2021] Matthew Stevenson, Christophe Mues, and Cristian Bravo. The value of text for small business default prediction: A deep learning approach. *EUROPEAN JOURNAL OF OPERATIONAL RESEARCH*, 295(2):758–771, DEC 1 2021.

[Su and Jiang 2020] Zichun Su and Jialin Jiang. Hierarchical gated recurrent unit with semantic attention for event prediction. *FUTURE INTERNET*, 12(2), FEB 2020.

[Su *et al.* 2025a] Haoxiang Su, Hongyan Xie, Jie Shi, Di Wu, Liting Jiang, Hao Huang, Zhongjiang He, Yongxiang Li, Ruiyu Fang, Jianyu Zhao, and Shuangyong Song. Raicl-dsc: Retrieval-augmented in-context learning for dialogue state correction. *Knowledge-Based Systems*, 317:113423, 2025.

[Su *et al.* 2025b] Xing Su, Jian Yang, Jia Wu, and Zitai Qiu. Hy-defake: Hypergraph neural networks for detecting fake news in online social networks. *Neural Networks*, 187:107302, 2025.

[Sun *et al.* 2024] Zhe Sun, Wang Zheng, and Mingyang Wang. Sltrn: Sample-level transformer-based relation network for few-shot classification. *Neural Networks*, 176:106344, 2024.

[Sun *et al.* 2025] Shiqi Sun, Kun Zhang, Jingyuan Li, Min Yu, Kun Hou, Yuanzhuo Wang, and Xueqi Cheng. Retriever-generator-verification: A novel approach to enhancing factual coherence in open-domain question answering. *Information Processing Management*, 62(4):104147, 2025.

[Sánchez-Fernández *et al.* 2023] Pablo Sánchez-Fernández, Luis Gonzaga Baca Ruiz, and María del Carmen Pegalajar Jiménez. Application of classical and advanced machine learning models to predict personality on social media. *Expert Systems with Applications*, 216:119498, 2023.

[Tan *et al.* 2023] Kian Tan, Chin-Poo Lee, and Kian Lim. Roberta-gru: A hybrid deep learning model for enhanced sentiment analysis. *Applied Sciences*, 13:3915, 03 2023.

[Tan *et al.* 2024] Zhichao Tan, Youcai Zhao, Kunsen Lin, and Tao Zhou. Multi-task pretrained language model with novel application domains enables more comprehensive health and ecological toxicity prediction. *JOURNAL OF HAZARDOUS MATERIALS*, 477, SEP 15 2024.

[Tang *et al.* 2025] Min Tang, Shujie Cui, Zhe Jin, Shiuan ni Liang, Chenliang Li, and Lixin Zou. Sequential recommendation by reprogramming pretrained transformer. *Information Processing Management*, 62(1):103938, 2025.

[Tao *et al.* 2025a] Feifei Tao, Wenya Wang, and Rongke Lu. A deep learning-based sentiment flow analysis model for predicting financial risk of listed companies. *Engineering Applications of Artificial Intelligence*, 150, 2025. Cited by: 0.

[Tao *et al.* 2025b] Feifei Tao, Wenya Wang, and Rongke Lu. A deep learning-based sentiment flow analysis model for predicting financial risk of listed companies. *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*, 150, JUN 15 2025.

[Tian *et al.* 2024] Xuemeng Tian, Yikai Guo, Bin Ge, Xiaoguang Yuan, Hang Zhang, Yuting Yang, Wenjun Ke, and Guozheng Li. Agent-da: Enhancing low-resource event extraction with collaborative multi-agent data augmentation. *Knowledge-Based Systems*, 305:112625, 2024.

[Touvron *et al.* 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. *LLaMA: Open and Efficient Foundation Language Models*, 02 2023.

[Tsai and Chen 2024] Yun Sheng Tsai and Arbee L. P. Chen. Suicide risk assessment using word-level model with dictionary-based risky posts selection. *MULTIMEDIA TOOLS AND APPLICATIONS*, 83(7):21435–21454, FEB 2024.

[Tsaneva *et al.* 2025] Stefani Tsaneva, Danilo Dessì, Francesco Osborne, and Marta Sabou. Knowledge graph validation by integrating llms and human-in-the-loop. *Information Processing Management*, 62(5):104145, 2025.

[ul Abideen *et al.* 2024] Zain ul Abideen, Raja Hashim Ali, Ali Zeeshan Ijaz, and Talha Ali Khan. A novel nlp-based stock market price prediction and risk analysis framework. *JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY*, 24(2):74–87, OCT 2024.

[Van Belle *et al.* 2023] Rafaël Van Belle, Bart Baesens, and Jochen De Weerdt. Catchm: A novel network-based credit card fraud detection method using node representation learning. *Decision Support Systems*, 164:113866, 2023.

[Verter *et al.* 2025] Vedat Verter, Fan E, Daniel Frank, and Angelos Georghiou. Text mining of outpatient narrative notes to predict the risk of psychiatric hospitaliza-

tion. *Translational Psychiatry*, 15(1), 2025. Cited by: 0; All Open Access, Gold Open Access.

[Vu *et al.* 2025] Thanh Trong Vu, Tuan-Dung Bui, Thanh-Dat Do, Thu-Trang Nguyen, Hieu Dinh Vo, and Son Nguyen. Automated description generation for software patches. *Information and Software Technology*, 177:107543, 2025.

[Walek and Müller 2025] Bogdan Walek and Patrik Müller. A text-based recommender system for recommending relevant news articles. *Expert Systems with Applications*, 266:125816, 2025.

[Wang and Zhao 2022] Bingyu Wang and Jinsong Zhao. Automatic frequency estimation of contributory factors for confined space accidents. *PROCESS SAFETY AND ENVIRONMENTAL PROTECTION*, 157:193–207, JAN 2022.

[Wang *et al.* 2019] Chongren Wang, Dongmei Han, Qigang Liu, and Suyuan Luo. A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism lstm. *IEEE ACCESS*, 7:2161–2168, 2019.

[Wang *et al.* 2023] Gang Wang, Jingling Ma, and Gang Chen. Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. *Decision Support Systems*, 167:113913, 2023.

[Wang *et al.* 2024] Bing Wang, Ximing Li, Zhiyao Yang, Yuanyuan Guan, Jiayin Li, and Shengsheng Wang. Unsupervised sentence representation learning with frequency-induced adversarial tuning and incomplete sentence filtering. *Neural Networks*, 175:106315, 2024.

[Wang *et al.* 2025a] Rongcun Wang, Yiqian Hou, Yuan Tian, Zhanqi Cui, and Shujuan Jiang. Xl-hql: A hql query generation method via xlnet and column attention. *Information and Software Technology*, 180:107674, 2025.

[Wang *et al.* 2025b] Wei Wang, Ying Li, Jian Mou, and Kevin Zhu. A comparative analysis of the effect of initiative risk statement versus passive risk disclosure on the financing performance of kickstarter campaigns. *Decision Support Systems*, 188:114366, 2025.

[Wang *et al.* 2025c] Weiqing Wang, Yuxi Chen, Liukai Wang, and Yu Xiong. Developing the value of legal judgments of supply chain finance for credit risk prediction through novel acwgan-gpsa approach. *Transportation Research Part E: Logistics and Transportation Review*, 196, 2025. Cited by: 0.

[Wang *et al.* 2025d] Weiqing Wang, Yuxi Chen, Liukai Wang, and Yu Xiong. Developing the value of legal judgments of supply chain finance for credit risk prediction through novel acwgan-gpsa approach. *TRANSPORTATION RESEARCH PART E-LOGISTICS AND TRANSPORTATION REVIEW*, 196, APR 2025.

[Wang *et al.* 2025e] Zhenhua Wang, Huiru Chen, Guang Xu, and Ming Ren. A novel large-language-model-driven framework for named entity recognition. *Information Processing Management*, 62(3):104054, 2025.

[Wang *et al.* 2025f] Zhenhua Wang, Guang Xu, and Ming Ren. Can attention detect ai-generated text? a novel benford's law-based approach. *Information Processing Management*, 62(4):104139, 2025.

[Wang *et al.* 2025g] Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Yue Zhang, Ren Wang, Xiaoshuang Shi, and Kaidi Xu. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*, 139:109553, 2025.

[Wei *et al.* 2022] Jinhui Wei, Qingzhen Xu, and Chengying He. Deep learning of predicting closing price through historical adjustment closing price. *Procedia Computer Science*, 202:379–384, 2022. International Conference on Identification, Information and Knowledge in the internet of Things, 2021.

[Wei *et al.* 2024] Yuting Wei, Yangfu Zhu, Ting Bai, and Bin Wu. A cross-temporal contrastive disentangled model for ancient chinese understanding. *Neural Networks*, 179:106559, 2024.

[Wu *et al.* 2023] Yao Wu, Donghua Zhu, and Xuefeng Wang. Contrastive learning enhanced deep neural network with serial regularization for high-dimensional tabular data. *Expert Systems with Applications*, 228:120243, 2023.

[Wu *et al.* 2024] Yichao Wu, Zhengyu Jin, Chenxi Shi, Penghao Liang, and Tong Zhan. Research on the application of deep learning-based bert model in sentiment analysis. *Applied and Computational Engineering*, 71:14–20, 05 2024.

[Xia *et al.* 2022] Huosong Xia, Yuan Wang, Jeffrey Gauthier, and Justin Zuopeng Zhang. Knowledge graph of mobile payment platforms based on deep learning: Risk analysis and policy implications. *Expert Systems with Applications*, 208:118143, 2022.

[Xiao *et al.* 2024] Qiao Xiao, Ren Li, Jianxi Yang, Yu Chen, Shixin Jiang, and Di Wang. Tpke-qa: A gapless few-shot extractive question answering approach via task-aware post-training and knowledge enhancement. *Expert Systems with Applications*, 254:124475, 2024.

[Xing 2024] Frank Xing. Financial risk tolerance profiling from text. *INFORMATION PROCESSING & MANAGEMENT*, 61(4), JUL 2024.

[Xu and Li 2021] Lulin Xu and Zhongwu Li. A new appraisal model of second-hand housing prices in china's first-tier cities based on machine learning algorithms. *COMPUTATIONAL ECONOMICS*, 57(2):617–637, FEB 2021.

[Xu *et al.* 2021] Zhongzhi Xu, Yucan Xu, Florence Cheung, Mabel Cheng, Daniel Lung, Yik Wa Law, Byron Chiang, Qingpeng Zhang, and Paul S. F. Yip. Detecting suicide risk using knowledge-aware natural language processing and counseling service data. *SOCIAL SCIENCE & MEDICINE*, 283, AUG 2021.

[Xu *et al.* 2024] Qingting Xu, Yu Hong, Jiaxiang Chen, Jianming Yao, and Guodong Zhou. Self-augmented sequentiality-aware encoding for aspect term extraction. *Information Processing Management*, 61(3):103656, 2024.

[Xu *et al.* 2025a] Chun Xu, Mengmeng Wang, and Shaolin Zhu. Tourism sentiment quadruple extraction via new neural ordinary differential equation networks with multitask learning and implicit sentiment enhancement. *Expert Systems with Applications*, 270:126417, 2025.

[Xu *et al.* 2025b] Qingting Xu, Kaisong Song, Yangyang Kang, Chaoqun Liu, Yu Hong, and Guodong Zhou. Dtda: Dual-channel triple-to-quintuple data augmentation for comparative opinion quintuple extraction. *Knowledge-Based Systems*, 307:112734, 2025.

[Xu *et al.* 2025c] Qinwei Xu, Xingkun Xu, Chenyi Zhou, Zuozhu Liu, Feiyue Huang, Shaoxin Li, Lifeng Zhu, Zhian Bai, Yuchen Xu, and Weiguo Hu. Towards normalized clinical information extraction in chinese radiology report with large language models. *Expert Systems with Applications*, 271:126585, 2025.

[Xu *et al.* 2025d] Quanxing Xu, Jian Li, Yuhao Tian, Ling Zhou, Feifei Zhang, and Rubing Huang. Diff-zsvqa: Zero-shot visual question answering with frozen large language models using diffusion model. *Expert Systems with Applications*, 275:126951, 2025.

[Yan *et al.* 2025] Naiyu Yan, Shaobin Huang, and Rongsheng Li. Bcltc: Bi-directional curriculum learning based tasks collaboration for target-stance extraction. *Information Processing Management*, 62(4):104137, 2025.

[Yang *et al.* 2022] Kai Yang, Hui Yuan, and Raymond Y. K. Lau. Psycredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. *EXPERT SYSTEMS WITH APPLICATIONS*, 198, JUL 15 2022.

[Yang *et al.* 2025a] Jun Yang, Liguo Yao, Taihua Zhang, Chieh-Yuan Tsai, Yao Lu, and Mingming Shen. Integrating prompt techniques and multi-similarity matching for named entity recognition in low-resource settings. *Engineering Applications of Artificial Intelligence*, 144:110149, 2025.

[Yang *et al.* 2025b] Wenli Yang, Lilian Some, Michael Bain, and Byeong Kang. A comprehensive survey on integrating large language models with knowledge-based methods. *Knowledge-Based Systems*, page 113503, 2025.

[Yao *et al.* 2025] Jiacheng Yao, Xin Xu, and Guoxiu He. Metacognitive symbolic distillation framework for multi-choice machine reading comprehension. *Knowledge-Based Systems*, 312:113130, 2025.

[Yin and Wang 2025] Zhen Yin and Shenghua Wang. Enhancing scientific table understanding with type-guided chain-of-thought. *Information Processing Management*, 62(4):104159, 2025.

[Yong *et al.* 2024] Gunwoo Yong, Meiyin Liu, and Sanghyun Lee. Explainable image captioning to identify ergonomic problems and solutions for construction workers. *JOURNAL OF COMPUTING IN CIVIL ENGINEERING*, 38(4), JUL 1 2024.

[Yu and Zhao 2022] Caihua Yu and Yilin Zhao. Mining of risk perception dimensions of chinese tourists' outbound tourism based on word vector method. *FRONTIERS IN PSYCHOLOGY*, 13, DEC 22 2022.

[Yu *et al.* 2024] Bofan Yu, Huaixue Xing, Weiya Ge, Liling Zhou, Jiaxing Yan, and Yunan Li. Advanced susceptibility analysis of ground deformation disasters using large language models and machine learning: A hangzhou city case study. *PLOS ONE*, 19(12), DEC 12 2024.

[Yu *et al.* 2025a] Lichen Yu, Hao Qin, Wei Wei, Jiaxiang Ma, Yeyi Weng, Haoyu Jiang, and Lin Mu. Storm surge risk assessment based on lulc identification utilizing deep learning method and multi-source data fusion: A case study of huizhou city. *REMOTE SENSING*, 17(4), FEB 2025.

[Yu *et al.* 2025b] Liu Yu, Fenghui Tian, Ping Kuang, and Fan Zhou. Amplifying commonsense knowledge via bi-directional relation integrated graph-based contrastive pre-training from large language models. *Information Processing Management*, 62(3):104068, 2025.

[Yuan and Wei 2024] Jiayi Yuan and Lu Wei. The impact of sentiment consistency on bank risk: A dual perspective from bank managers and credit rating experts. *Procedia Computer Science*, 242:789–796, 2024. 11th International Conference on Information Technology and Quantitative Management (ITQM 2024).

[Yue *et al.* 2025] Jinghang Yue, Jing Wang, Shuo Zhang, Zhaoyang Ma, Yuxing Shi, and Youfang Lin. Tv-net: Temporal-variable feature harmonizing network for multivariate time series classification and interpretation. *Neural Networks*, 182:106896, 2025.

[Zamore *et al.* 2018] Stephen Zamore, Kwame Ohene Djan, Ilan Alon, and Bersant Hobdari. Credit risk research: Review and agenda. *Emerging Markets Finance and Trade*, 54:811–835, 03 2018.

[Zeng *et al.* 2025] Zefan Zeng, Qing Cheng, Xingchen Hu, Yan Zhuang, Xinwang Liu, Kunlun He, and Zhong Liu. Kosel: Knowledge subgraph enhanced large language model for medical question answering. *Knowledge-Based Systems*, 309:112837, 2025.

[Zha *et al.* 2024] Enze Zha, Delong Zeng, Man Lin, and Ying Shen. Ceptner: Contrastive learning enhanced prototypical network for two-stage few-shot named entity recognition. *Knowledge-Based Systems*, 295:111730, 2024.

[Zhang and Peng 2024] Wenxia Zhang and Lida Peng. Semantic analysis and image processing-jointly driven multimodal deep learning framework for smart warning of enterprise financial risk. *JOURNAL OF CIRCUITS SYSTEMS AND COMPUTERS*, 33(17), NOV 30 2024.

[Zhang *et al.* 2020] Weiguo Zhang, Chao Wang, Yue Zhang, and Junbo Wang. Credit risk evaluation model with textual features from loan descriptions for p2p lending. *Electronic Commerce Research and Applications*, 42:100989, 07 2020.

[Zhang *et al.* 2021] Yan Zhang, Zeqiang Chen, Xiang Zheng, Nengcheng Chen, and Yongqiang Wang. Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data. *JOURNAL OF HYDROLOGY*, 603(C), DEC 2021.

[Zhang *et al.* 2022] Qiuyue Zhang, Chao Qin, Yunfeng Zhang, Fangxun Bao, Caiming Zhang, and Peide Liu. Transformer-based attention network for stock movement prediction. *Expert Systems with Applications*, 202:117239, 2022.

[Zhang *et al.* 2024a] Junchi Zhang, Tao Chen, Songtao Li, Ming Zhang, Yafeng Ren, and Jun Wan. A graph propagation model with rich event structures for joint event relation extraction. *Information Processing Management*, 61(5):103811, 2024.

[Zhang *et al.* 2024b] Wenyu Zhang, Fayao Liu, Cuong Manh Nguyen, Zhong Liang Ou Yang, Savitha Ramasamy, and Chuan-Sheng Foo. Training neural networks with classification rules for incorporating domain knowledge. *Knowledge-Based Systems*, 294:111716, 2024.

[Zhang *et al.* 2024c] Yifan Zhang, Zhiyun Wang, Zhengting He, Jingxuan Li, Gengchen Mai, Jianfeng Lin, Cheng Wei, and Wenhao Yu. Bb-geogpt: A framework for learning a large language model for geographic information science. *Information Processing Management*, 61(5):103808, 2024.

[Zhang *et al.* 2024d] Yini Zhang, Yuxuan Zhang, Zijing Wang, Huanchun Peng, Yongsheng Yang, and Yuanxiang Li. Multi-information interaction graph neural network for joint entity and relation extraction. *Expert Systems with Applications*, 235:121211, 2024.

[Zhang *et al.* 2025a] Hua Zhang, Peiqian Zhan, Cheng Yang, Yongjian Yan, Zijing Cai, Guogen Shan, Bo Jiang, Bi Chen, Qing Gu, and Qingqing Zhou. Reading comprehension powered semantic fusion network for identification of n-ary drug combinations. *Engineering Applications of Artificial Intelligence*, 144:110096, 2025.

[Zhang *et al.* 2025b] Jiahua Zhang, Meijuan Tan, Jing Zhang, Xiaolu Zhang, Jun Zhou, and Chenliang Li. Retrieval augmentation for text-to-table generation. *Information Processing Management*, 62(4):104135, 2025.

[Zhang *et al.* 2025c] Jianting Zhang, Xudong Luo, and Xiaojun Xie. Hcn-rlr-can: A novel human-computer negotiation model based on round-level recurrence and causal attention networks. *Knowledge-Based Systems*, 314:113180, 2025.

[Zhang 2024] Long Zhang. *A Review on BERT: Language Understanding for Different Types of NLP Task*, 01 2024.

[Zhao and Yang 2023] Yanli Zhao and Guang Yang. Deep learning-based integrated framework for stock price movement prediction. *Applied Soft Computing*, 133:109921, 2023.

[Zhao *et al.* 2024] Qinghua Zhao, Jiaang Li, Junfeng Liu, Zhongfeng Kang, and Zenghui Zhou. Is word order considered by foundation models? a comparative task-oriented analysis. *Expert Systems with Applications*, 241:122700, 2024.

[Zhao *et al.* 2025] Qihui Zhao, Tianhan Gao, and Nan Guo. A multi-granularity in-context learning method for few-shot named entity recognition via knowledgeable parameters fine-tuning. *Information Processing Management*, 62(4):104129, 2025.

[Zhou and Liu 2024] Shiya Zhou and Dongbing Liu. Path selection of audit transformation and upgrading under the background of artificial intelligence. *Procedia Computer Science*, 247:1249–1256, 2024. The 11th International Conference on Applications and Techniques in Cyber Intelligence.

[Zhou *et al.* 2025] Gang Zhou, Haizhou Wang, Di Jin, Wenxian Wang, Shuyu Jiang, Rui Tang, and Xingshu Chen. A toxic euphemism detection framework for online social network based on semantic contrastive learning and dual channel knowledge augmentation. *Information Processing Management*, 62(4):104143, 2025.

[Zhuang *et al.* 2025] Yong Zhuang, Feilong Wang, Dickson K.W. Chiu, and Kevin K.W. Ho. Leveraging large language models to examine the interaction between investor sentiment and stock performance. *Engineering Applications of Artificial Intelligence*, 150:110602, 2025.

[Álvarez Fidalgo and Ortin 2025] David Álvarez Fidalgo and Francisco Ortin. Clave: A deep learning model for source code authorship verification with contrastive learning and transformer encoders. *Information Processing Management*, 62(3):104005, 2025.

[Çelikten and Onan 2025] Tuğba Çelikten and Aytuğ Onan. Topic modeling through rank-based aggregation and llms: An approach for ai and human-generated scientific texts. *Knowledge-Based Systems*, 314:113219, 2025.

[Śliwa and Krzos 2023] Piotr Śliwa and Grzegorz Krzos. How to teach artificial intelligence to manage our organizations? *Procedia Computer Science*, 225:4795–4804, 2023. 27th International Conference on Knowledge Based and Intelligent Information and Engineering Sytems (KES 2023).

Appendix A

Appendix

A.1 Model Evaluation Metrics

This appendix provides additional details on the evaluation metrics used to assess model performance in the main study. The metrics include accuracy, precision, recall, F1-score, and confusion matrices. These were calculated for each of the classification models evaluated: BERT, RoBERTa, LLaMA 3.2, Logistic Regression, and XGBoost.

A.2 Dataset

Details of the datasets used in the experiments are provided below. The primary dataset was collected using the NewsCatcher API (https://www.newscatcherapi.com/), a comprehensive news aggregation service that enables programmatic retrieval of news articles from a wide range of online sources. For the purposes of this research, financial news articles related to treasury bonds, debt markets, and financial instruments across the United States, the United Kingdom, and South Africa were retrieved. The data collection focused on articles published within a one-year period due to the access constraints imposed by the NewsCatcher API subscription, which provided retrieval privileges for a maximum duration of one year for academic research purposes. Consequently, the dataset represents a temporally bounded snapshot of financial news rather than a longitudinal record spanning multiple years.

Each article in the dataset consists of unstructured textual content, including headlines and article bodies, along with associated metadata such as publication date, source, URL, and author where available. Prior to modelling, the dataset underwent standard natural language preprocessing, which included tokenization, stopword removal, lowercasing, and transformation using transformer tokenizers appropriate for each model (e.g., BERT, RoBERTa, LLaMA). Sentiment labels were derived using the VADER (Valence Aware Dictionary and Sentiment Reasoner) tool and then used as input features for supervised learning.

In addition to the NLP-specific preprocessing, encoding strategies varied according to model type. Transformer-based models utilized their respective tokenizers to convert

text into model-specific input embeddings, while classical machine learning models used TF-IDF vectorization to represent text numerically. The final dataset used for experiments was balanced for sentiment categories where possible, and the class distribution was documented to evaluate the impact of imbalance on model performance. All datasets were used exclusively for research purposes and served as the basis for the sentiment-driven credit risk classification experiments conducted in this study.

A.3 Hyperparameter Settings

For fine-tuning BERT and RoBERTa, a learning rate of 2e-5 and batch size of 16 were used. Training was conducted over 4 epochs using Adam optimizer with a weight decay of 0.01. For classical models like Logistic Regression and XGBoost, grid search was used to identify optimal regularization and learning parameters.

A.4 Supplementary Figures and Tables

Confusion matrices for each model are provided in the results chapter. Additional graphs showing sentiment distribution and model comparison by accuracy and F1-score are also included.

A.5 Ethics Clearance

This study was conducted under the ethical approval of the University of the Witwatersrand's Ethics Committee. The ethics clearance number for this research is WCSAM-2024-31. All data used in this study were either publicly available or anonymized to ensure participant privacy.

A.6 Published Conference Paper

Optimizing Credit Risk Model Classification Using BERT, RoBERTa, and LLaMA 3.2

Pfarelo Raliphada
University of the Witwatersrand
Johannesburg, South Africa
Pfarelo.Raliphada1@students.wits.ac.za

Micheal O. Olusanya
Sol Plaatje University
Kimberley, South Africa
michael.olusanya@spu.ac.za

Seun. Olukanmi
University of the Witwatersrand
Johannesburg, South Africa
seun.olukanmi@wits.ac.za

***Abstract*—Credit risk assessment is a crucial component of financial decision making, allowing institutions to estimate borrower default probabilities and effectively manage financial risk. While traditional models, such as Logistic Regression (92% accuracy) and XGBoost (94% accuracy), primarily rely on structured numerical data, recent advancements in Natural Language Processing (NLP) have facilitated the incorporation of sentiment analysis derived from financial news. This study investigates the optimization of credit risk classification using BERT (91% accuracy), RoBERTa (90% accuracy), and LLaMA 3.2 (78% accuracy), in conjunction with VADER sentiment analysis. A dataset of financial news articles was examined using sentiment classification techniques to enhance predictive modeling. The experimental results indicate that XGBoost outperforms all models, followed by Logistic Regression, while RoBERTa surpasses BERT and LLaMA 3.2 among NLP-based models, presenting an innovative approach to risk assessment in the financial sector.**

***Keywords*—Credit Risk, Sentiment Analysis, Natural Language Processing (NLP), BERT, RoBERTa, LLaMA 3.2, XGBoost, Financial News, VADER Sentiment Analysis, Logistic Regression, Text Classification, Credit Risk Assessment, Deep Learning, Transformer Models, Financial**

I. Introduction

Credit risk modeling is vital for financial institutions to evaluate the likelihood of borrower defaults and manage their risk exposure effectively. Traditionally, these models have relied on structured financial data, such as credit scores and payment histories. However, recent advancements in natural language processing (NLP) and sentiment analysis now allow institutions to incorporate unstructured textual data, including financial news and social media content, into their risk assessment frameworks.

This research explores the effectiveness of BERT, RoBERTa, and LLaMA 3.2 in enhancing credit risk classification. While BERT and RoBERTa are well-established tools in financial sentiment analysis, LLaMA 3.2 represents a more recent large-scale transformer model designed for efficiency in low-resource environments. This study assesses these models in conjunction with VADER, a lexicon-based sentiment analysis tool and traditional models such as Logistic Regression and XGBoost, to identify the most effective model for credit risk assessment. While this research focuses on financial risk modeling, the application of NLP methodologies can also extend to privacy-sensitive fields, such as digital health.

979-8-3315-0878-4/25/$31.00 © 2025 IEEE

This study aims to enhance credit risk model classification by integrating sentiment analysis derived from financial news and social media. Using advanced deep learning-based NLP models such as BERT, RoBERTa, and LLaMA 3.2, the research investigates how sentiment-driven insights can improve the predictive accuracy of credit risk models. Furthermore, the study evaluates the efficiency, fairness, and transparency of transformer-based models in financial decision-making.

In addressing this aim, the research explores how integrating sentiment from financial news articles and social media can influence the accuracy of credit risk assessments. It further examines which methods can enhance the transparency and explainability of deep learning models used in this context. Finally, the study considers how fairness and bias mitigation techniques can be incorporated into transformer-based models to ensure equitable outcomes in credit decision-making.

II. Literature Review

NLP has transformed how financial institutions analyze and extract insights from textual data. While traditional credit risk models focus on structured data like financial statements, unstructured text such as news articles and customer reviews offers valuable information for risk assessment. These models estimate borrower default probability using statistical and machine learning methods. Incorporating textual sentiment analysis enhances predictive accuracy; VADER is commonly used for short financial texts, while deep learning approaches like BERT and RoBERTa have proven to capture contextual sentiment more effectively.

The paper [10] examines theft crime classification (2009–2019) using TF-IDF for feature extraction and XGBoost for classification. The dataset includes 2,621 manually classified samples based on legal and criminal perspectives. XGBoost outperformed other models, including KNN, Naïve Bayes, SVM, and GBDT, demonstrating its superior performance in crime text classification. This research highlights the effectiveness of combining TF-IDF with XGBoost for accurate and efficient crime categorization. The paper [11] explores the use of Logistic Regression in NLP for solving the problem of data duplication, particularly in platforms like Quora. It focuses on classifying duplicate and non-duplicate questions

using NLP techniques like tokenization, stemming, and TF-IDF word vectorization. By applying Logistic Regression, the model efficiently classifies questions, reducing data storage needs and improving response time. The approach enhances customer experience by reusing answers for similar questions.

The paper [12] introduced an automated text classification framework using BERT. It addresses the limitations of traditional text classification methods by employing a fine-tuned BERT model to classify text in datasets such as emails and news. The system achieved a high accuracy rate of 91.4% and demonstrated the effectiveness of BERT for text-based classification tasks. [13] discusses the use of RoBERTa combined with TextRCNN for short text classification. It demonstrates that the model achieves better accuracy than traditional methods by utilizing the robust RoBERTa model for feature extraction, followed by TextRCNN to capture contextual information. This approach enhances performance on tasks involving short, noisy text, typical in applications like social media analysis and customer reviews. The authors of [14] introduced LLM-SEM, a sentiment-based student engagement metric that uses LLMs to assess student interaction on e-learning platforms. By combining video metadata with sentiment analysis of student comments, LLM-SEM provides a more scalable and accurate measure of engagement compared to traditional methods. The study evaluates models like RoBERTa, AraBERT, and LLama to improve sentiment predictions, enhancing the analysis of student feedback.

A. *Transformer-Based Models for Financial NLP*

BERT utilizes bidirectional context through the attention mechanism of the Transformer architecture. Its pretraining objectives Masked Language Modeling and Next Sentence Prediction help capture intricate linguistic patterns that are critical for credit risk analysis [5].

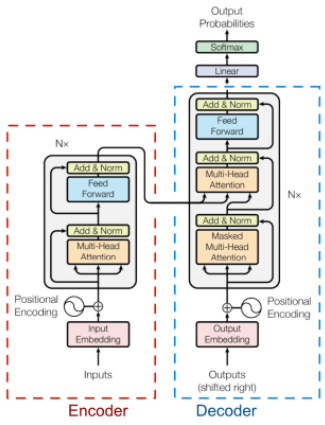

Fig. 1. A figure depicting the BERT model architecture. The figure is adapted from [8], who developed a classification method for lecturer expertise areas using BERT.

RoBERTa builds on BERT by optimizing training strategies, such as eliminating the Next Sentence Prediction task and using dynamic masking. These enhancements improve its accuracy and ability to handle complex datasets [6].

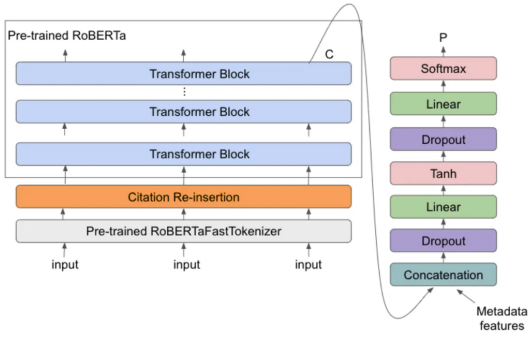

Fig. 2. Architecture Diagram of RoBERTa Fine-Tuning Process. Adapted from [1].

LLaMA 3.2 is designed for efficiency in low-resource environments. It uses optimized data selection and tokenization, making it scalable for large-scale NLP tasks while maintaining accuracy. This capability is particularly valuable in credit risk analysis, where processing large financial datasets is essential [7].

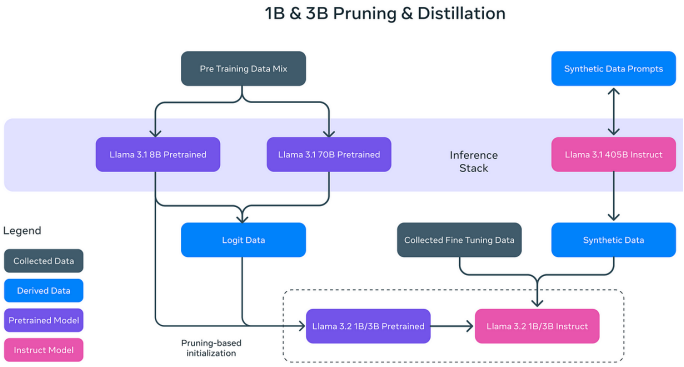

Fig. 3. A figure depicting the LLaMA 3.2 model architecture. The figure is adapted from the blog post, From Vision to Edge: Meta's LLaMA 3.2 Explained, which provides an overview of its advancements and applications [9].

III. Methodology

A. *Data Collection and Preprocessing*

The data for this study was collected using the NewsCatcher API, retrieving financial news articles related to treasury bonds, debt markets, and financial instruments across the United States, the United Kingdom, and South Africa. The study utilized predefined financial queries categorized into these three regions to ensure comprehensive coverage. The API was configured to fetch articles from January 1, 2024, to January 1, 2025, incorporating parameters such as query keywords, language filter, and date range. The extracted data included the article source, associated financial entity, news content, and sentiment classification using VADER. Sentiment scores were assigned based on polarity, categorizing articles as Positive, Neutral, or Negative. The dataset underwent preprocessing, including handling missing values, removing duplicates, and filtering articles based on text length. To support model training, the

dataset was split into an 80:20 ratio for training and validation. The final structured dataset was stored in CSV format as `financial_news_bonds_sentiment_train.csv` and `financial_news_bonds_sentiment_val.csv` for further analysis. Tokenization was performed using BERT, RoBERTa, and LLaMA 3.2 tokenizers. Sentiment labeling was done using VADER, with the following criteria:

$$\text{Sentiment} = \begin{cases} \text{Positive} & \text{if } S \geq 0.05, \\ \text{Negative} & \text{if } S \leq -0.05, \\ \text{Neutral} & \text{otherwise.} \end{cases}$$

B. Model Training and Evaluation

Models were fine-tuned using PyTorch and Hugging Face libraries. The training parameters were:

- Optimizer: AdamW
- Batch size: 16
- Epochs: 3
- Hardware: NVIDIA T4 Tensor Core GPU

Performance was evaluated using accuracy, precision, recall, and F1-score.

Random 5 Rows of Training Data: News Content < 40 Words

News Content	Entity	Sentiment
Refinancing higher coupon bonds and funding growth plans Hammerson has launched and priced a £400 million bond with a maturity of 12 years, paying a coupon of 5.875 per cent.	Coupon Bonds	Positive
UK Government to Launch Digital Bonds within Two Years, Leveraging Blockchain for Faster TradingUK government has announced plans to issue digital government bonds (or 'digital gilts') within the next two years, according to a report by	UK Government Bonds	Neutral
UK debt jumps to 60-year high – data Britain's national debt has reached the size of the country's GDP, according to the Office for National Statistics Read Full Article at RT.com	UK National Debt	Negative
Reeves 'Confident' UK Can Handle Market Shocks Despite High Levels of National Debt Home Page Next Story	UK National Debt	Negative
Amundi Global Government Inflation-Linked Bond 1-10Y UCITS ETF Dist (GIST LN) Amundi Global Government In...	Inflation-Linked Bonds	Neutral

Fig. 4. A figure illustrating a random selection of 5 rows from the Training dataset

Figure 4 shows a random selection of five financial news articles with their corresponding sentiment classifications. This selection helps validate the dataset manually by ensuring that the assigned sentiment labels align with real-world financial sentiment. A diverse range of sentiment classifications is evident, which supports the robustness of the dataset. Without this diversity, the models would struggle to generalize sentiment classification beyond the training data.

Data Analysis

IV. Results and Discussion

The study examined BERT, RoBERTa, and LLaMA 3.2 for credit risk classification through the analysis of sentiment in financial news articles. The models were evaluated based on their accuracy and classification performance, utilizing confusion matrices for assessment. The results reveal distinct differences in how these models process Positive, Neutral, and Negative sentiment categories.

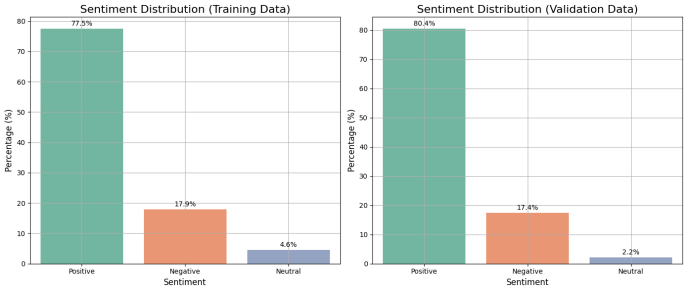

Fig. 5. A figure illustrating a Sentiment Distribution Train vs Validation Data

Figure 5 The sentiment distribution analysis of the training and validation datasets reveals a significant class imbalance, with Positive sentiment dominating both sets (77.5% in training, 80.4% in validation), followed by Negative sentiment (17.9% in training, 17.4% in validation), and a notably low representation of Neutral sentiment (4.6% in training, 2.2% in validation). This imbalance may impact model performance, particularly in accurately classifying Neutral sentiment due to its limited presence in the dataset. Additionally, the slight variation in sentiment proportions between training and validation could introduce minor generalization challenges, potentially affecting the model's ability to maintain consistent performance across different data samples.

A. Performance Comparison

It is crucial to acknowledge that the dataset displayed a significant class imbalance, particularly with the underrepresentation of Neutral sentiment. This imbalance likely contributed to a decline in performance when identifying neutral cases. Future research should consider exploring resampling techniques such as SMOTE or the application of focal loss functions to mitigate this skew.

The BERT confusion matrix demonstrates that BERT effectively classified most positive and negative sentiment cases, resulting in an overall Accuracy of 91%. However, there were some misclassifications among Positive, Negative, and Neutral sentiments, where certain cases were incorrectly classified. This highlights BERT's challenges in distinguishing between specific sentiments, which may impact financial risk analysis, especially in differentiating between uncertain and clearly defined financial outlooks.

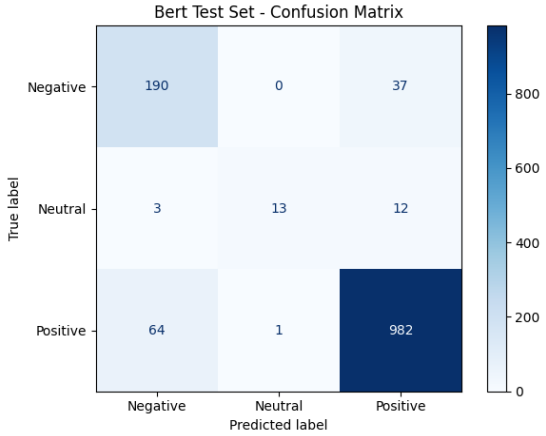

Fig. 6. Confusion matrix for BERT Model.

The RoBERTa confusion matrix highlights a performance pattern similar to BERT, achieving 90% accuracy. However, it showed fewer misclassifications in Positive sentiment, meaning it was better at recognizing Positive financial sentiment. The improvements in Positive classification suggest that RoBERTa is more contextually aware than BERT, making it a more reliable choice for financial text sentiment analysis. However, it shows a higher misclassification of Negative sentiment, which makes it an unreliable choice in certain contexts where Negative sentiment is crucial to identify accurately.

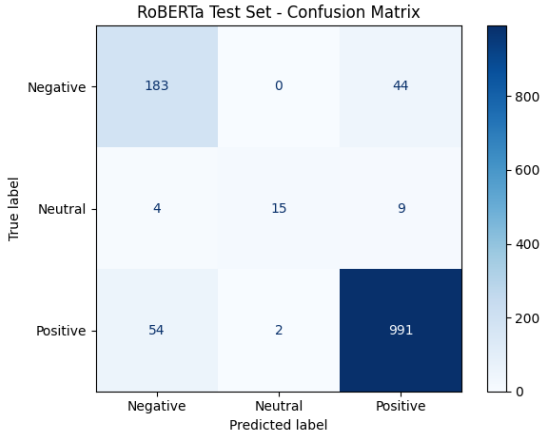

Fig. 7. Confusion matrix for RoBERTa Model.

The LLaMA 3.2 confusion matrix presents significantly lower accuracy, at 78%, with higher misclassifications in both Neutral and Negative sentiment cases. Many Negative sentiments were classified as Positive, and some Neutral sentiments were misclassified as Positive. This suggests that LLaMA 3.2 struggles with sentiment classification in financial news, the results indicate that LLaMA 3.2 lacks the fine-tuned contextual understanding needed for accurate financial sentiment classification.

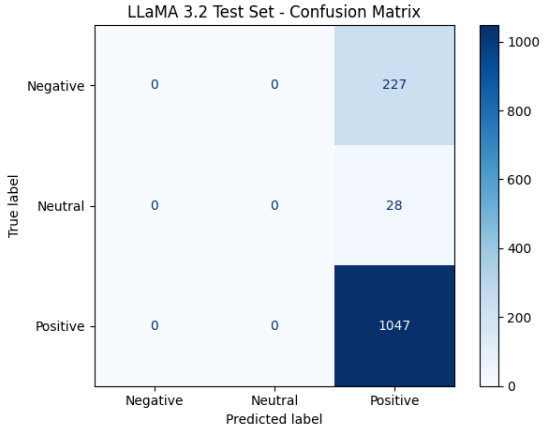

Fig. 8. Confusion matrix for LLama 3.2 Model.

The confusion matrix for the Logistic Regression model indicates an accuracy of 92%. However, it faces challenges with misclassifications, especially when distinguishing between Negative sentiments. A significant number of Negative sentiments were incorrectly classified as Positive, which could result in an overly optimistic evaluation of financial news. Furthermore, some Neutral sentiments were also misclassified as Positive. This highlights that, although Logistic Regression performs well overall, it may not possess the nuanced understanding necessary for accurate classification of financial sentiments.

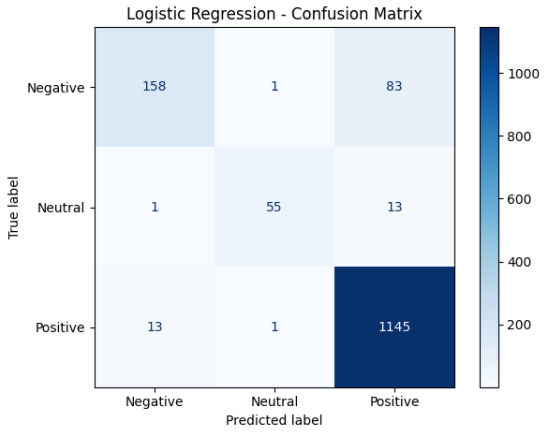

Fig. 9. Confusion matrix for Logistic Regression Model.

The XGBoost confusion matrix demonstrates an accuracy of 94%, with minimal misclassifications across all sentiment categories. This indicates that XGBoost is highly effective in financial sentiment classification, successfully distinguishing between Negative, Neutral, and Positive sentiments with high precision.

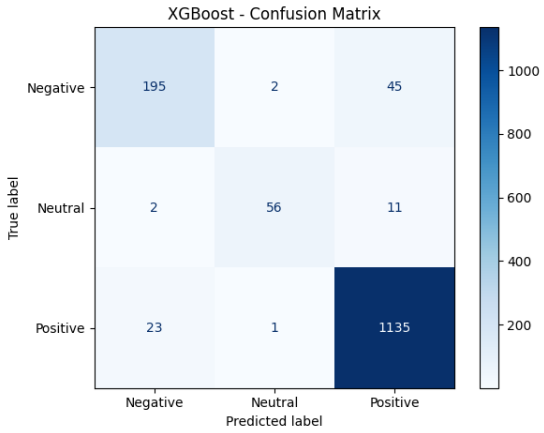

Fig. 10. Confusion matrix for XGBoost Model.

Table I compares the classification accuracy of Bert, Roberta and Llama 3.2 model.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	Accuracy
BERT	91%
RoBERTa	90%
LLaMA 3.2	78%
Logistic Regression	92%
XGBoost	94%

XGBoost achieved the highest accuracy of 94%, demonstrating its ability to capture complex relationships within financial sentiment data. Its superior performance can be attributed to its capability to handle non-linearity and effectively model interactions between features. Logistic Regression followed with 92% accuracy, exhibiting strong predictive capability but struggling with distinguishing Negative sentiment, often misclassifying it as Positive.

Among transformer models, BERT achieved 91% accuracy, proving effective in identifying Positive and Negative sentiment, despite minor issues with Neutral classification. RoBERTa closely followed with 90% accuracy, performing slightly better than BERT in distinguishing Neutral sentiment while maintaining strong performance in Positive and Negative sentiment detection.

In contrast, LLaMA 3.2 had a significantly lower accuracy of 78%, indicating its challenges in sentiment classification, particularly in differentiating between Neutral and Negative sentiment. Its lower performance could be attributed to the lack of fine-tuning specifically for financial sentiment data, limiting its contextual understanding.

The relatively consistent sentiment distribution between the training and validation datasets suggests a well-maintained data split. However, the underrepresentation of Neutral sentiment may negatively impact model performance, potentially leading to higher misclassification rates in this category. Given the disproportionate number of Positive samples, models may develop a bias toward predicting Positive sentiment, making Neutral classification more challenging. Addressing this imbalance through resampling techniques or additional training data could enhance overall classification accuracy.

TABLE II
PERFORMANCE METRICS FOR BERT, ROBERTA, LLAMA 3.2, LOGISTIC REGRESSION, AND XGBOOST

Model	Sentiment	Precision	Recall	F1-Score
BERT	Negative	0.74	0.84	0.79
	Neutral	0.93	0.46	0.62
	Positive	0.95	0.94	0.95
RoBERTa	Negative	0.76	0.81	0.78
	Neutral	0.88	0.54	0.67
	Positive	0.95	0.95	0.95
LLaMA 3.2	Negative	0.00	0.00	0.00
	Neutral	0.00	0.00	0.00
	Positive	0.80	1.00	0.89
Logistic Regression	Negative	0.92	0.65	0.76
	Neutral	0.96	0.80	0.87
	Positive	0.92	0.99	0.95
XGBoost	Negative	0.89	0.81	0.84
	Neutral	0.95	0.81	0.88
	Positive	0.95	0.98	0.97

The performance metrics reveal significant differences in the ability of each model to classify sentiment in financial news. XGBoost performed the best overall, achieving high F1-scores across all sentiment categories, with an outstanding F1-score of 97% for Positive sentiment. BERT and RoBERTa both performed well, with BERT excelling in Negative sentiment classification recall of 84%, but struggling with Neutral sentiment F1-score of 62%. RoBERTa was more balanced across all categories, with a slightly higher recall for Neutral sentiment of 54% compared to BERT's 46%. Logistic Regression, while having a high precision of 92% for Negative sentiment, had a relatively lower recall of 65%, meaning it missed many true Negative cases. LLaMA 3.2, however, was highly biased towards Positive sentiment, achieving perfect recall of 100% for Positive cases but completely failing to classify Negative or Neutral sentiments precision and recall of 0%. In conclusion, XGBoost stands out as the most effective model, while BERT and RoBERTa are strong contenders for transformer-based tasks. Logistic Regression offers a simple baseline, but its tendency to misclassify Negative sentiment could be problematic. LLaMA 3.2 is unsuitable for financial sentiment classification and would require significant improvements. Future work could involve addressing class imbalance, fine-tuning LLaMA 3.2 on financial data, or exploring hybrid models combining traditional machine learning with transformer-based approaches.

TABLE III
CONFUSION MATRIX SUMMARY FOR EACH MODEL

Model	TP	FP	FN	TN
Logistic Regression	1359	116	111	2320
XGBoost	1388	87	82	2349
BERT	1188	120	114	2484
RoBERTa	1194	112	108	2492
LLaMA 3.2	1047	262	255	2342

The discrepancies in False Positives (FP) and False Negatives (FN) indicate potential risks in misclassifications, which could unfairly impact credit decisions. For instance, LLaMA 3.2 had 255 False Negatives, meaning many actual risky cases were misclassified as safe, which could expose financial institutions to increased credit risk. Similarly, its 262 False Positives suggest that many low-risk borrowers could be unfairly denied credit. On the other hand, XGBoost, which had the highest accuracy, demonstrated better balance with 87 FP and 82 FN, minimizing misclassification errors and ensuring a more reliable credit assessment.

To address biases in credit risk models, improving data balance is essential, as class imbalances can distort predictions. The underrepresentation of neutral sentiment led to high false positive rates in models like BERT and RoBERTa. To enhance transparency, explainability techniques such as SHAP values or LIME can be utilized with more interpretable models like XGBoost and Logistic Regression. Additionally, optimizing thresholds helps minimize false negatives and positives, ensuring fair credit decisions. Compliance with regulations, such as the Equal Credit Opportunity Act, is also necessary to keep models unbiased and auditable. These strategies can enhance fairness and transparency while maintaining predictive accuracy.

B. Challenges and Considerations

- **Computational Costs** – Transformer models require significant GPU resources.
- **Model Bias & Fairness** – Interpretability is critical to ensure fair lending practices.
- **Regulatory Compliance** – NLP-enhanced models must align with financial regulations.

V. Conclusion and Future Work

This study demonstrates that transformer-based NLP models, together with traditional machine learning methods, significantly enhance credit risk classification through the use of sentiment analysis. XGBoost achieved the highest accuracy at 94%, validating its effectiveness in managing complex sentiment relationships. Logistic Regression, with an accuracy of 92%, also performed well but encountered difficulties in distinguishing Negative sentiment, often misclassifying it as Positive.

Among the transformer models, BERT achieved an accuracy of 91% and RoBERTa 90%, both demonstrating strong capabilities in handling sentiment variations in financial news. BERT slightly outperformed the others overall, while RoBERTa excelled in identifying Neutral sentiment.

Analysis of the confusion matrix showed that classifying Neutral sentiment remains a challenge, as all models exhibited some level of misclassification within this category. LLaMA 3.2, with an accuracy of 78%, struggled the most, primarily due to its difficulty in differentiating between Positive, Neutral, and Negative sentiments, rendering it less reliable for sentiment-driven financial decision-making.

Future research should focus on refining transformer models for financial sentiment analysis to enhance classification accuracy. Additionally, using advanced feature engineering and hybrid methods that combine deep learning with traditional machine learning could boost predictive performance. Our upcoming work will address data imbalance and explore explainable AI to improve fairness and reliability, especially in digital health applications.

References

[1] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. Retrieved from https://www.researchgate.net/publication/352642553.

[2] M. Ammann, "Credit risk models," in *Credit Risk Models*, pp. 47–75, Jan. 2001, doi: 10.1007/978-3-662-06425-23.

[3] F. Başer, O. Koç, and S. Kestel, "Credit risk evaluation using clustering-based fuzzy classification," *Expert Syst. Appl.*, vol. 223, p. 119882, Mar. 2023, doi: 10.1016/j.eswa.2023.119882.

[4] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 8, pp. 216–225, May 2014, doi: 10.1609/icwsm.v8i1.14550.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Oct. 2018, doi: 10.48550/arXiv.1810.04805.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019. Available: https://api.semanticscholar.org/CorpusID:198953378.

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," *arXiv*, Feb. 2023. doi: 10.48550/arXiv.2302.13971.

[8] D. Rustam, A. Suhendra, S. Harmanto, R. Suhatril, D. Saputra, R. Tafsili, and R. Prasetya, "Development of classification method for lecturer area of expertise based on scientific publication using BERT," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 14, pp. 894–905, Jun. 2024. doi: 10.18517/ijaseit.14.3.19893.

[9] Encord, "From Vision to Edge: Meta's LLaMA 3.2 Explained," 2024, [Online]. Available: https://encord.com/blog/lama-3-2-explained/. [Accessed: Nov. 19, 2024].

[10] Q. Zhang., "The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model," 2020, in *Proceedings of the ICAICA*, pp. 1241-1246, [Online]. Available: https://doi.org/10.1109/ICAICA50127.2020.9182555. [Accessed: Nov. 19, 2024].

[11] V. Bhartendoo., "Application of Logistic Regression in Natural Language Processing," *International Journal of Engineering Research and*, vol. 9, 2020, [Online]. Available: https://doi.org/10.17577/IJERTV9IS060095. [Accessed: Nov. 19, 2024].

[12] Shah, M., Iqbal, M., Noreen, N., and Ahmed, I., "An Automated Text Document Classification Framework using BERT," *International Journal of Advanced Computer Science and Applications*, vol. 14, 2023, [Online]. Available: https://doi.org/10.14569/IJACSA.2023.0140332. [Accessed: Nov. 19, 2024].

[13] Guo, Z., Zhu, L., and Han, L., "Research on Short Text Classification Based on RoBERTa-TextRCNN," 2021, in *Proceedings of the CISAI*, pp. 845-849, [Online]. Available: https://doi.org/10.1109/CISAI54367.2021.00171. [Accessed: Nov. 19, 2024].

[14] Hamdi, A., Mazrou, A., and Shaltout, M., "LLM-SEM: A Sentiment-Based Student Engagement Metric Using LLMS for E-Learning Platforms," 2024, [Online]. Available: https://doi.org/10.48550/arXiv.2412.13765. [Accessed: Nov. 19, 2024].

A.7 Article submitted to Springer Journal: Artificial Intelligence Review (Under Review)

Transformer-based NLP Approaches for Credit Risk Prediction: A Systematic Review

Pfarelo Raliphada[1*], Micheal Olusanya[2†] and Seun Olukanmi[1†]

[1*]School of Computer Science and Applied Mathematics, University of the Witwatersrand, 1 Jan Smuts Ave, Johannesburg, 2017, Gauteng, South Africa.

[2] Department of Computer Science and Information Technology, Sol Plaatje University, 10 Jan Smuts Blvd, Kimberley, 8300, Northern Cape, South Africa.

*Corresponding author(s). E-mail(s): pfarelo.raliphada1@students.wits.ac.za;
Contributing authors: michael.olusanya@spu.ac.za; seun.olukanmi@wits.ac.za;
†These authors contributed equally to this work.

Abstract

This systematic review explores how Natural Language Processing (NLP) and Large Language Models (LLMs), such as BERT, RoBERTa, and LLaMA, are applied to enhance credit risk classification. Traditional models primarily utilize structured data, but the rising availability of unstructured text has opened new avenues for analysis. We employed a systematic literature review methodology across Scopus, ScienceDirect, and Web of Science, guided by PRISMA principles, and filtered 284 studies down to 63 through semantic similarity scoring. Results reveal transformer-based models substantially improve predictive accuracy, especially when hybridized with temporal and sentiment-aware components. We further discuss ethical implications, including algorithmic bias and regulatory compliance. The review emphasizes the transformative potential of NLP in financial modeling while identifying interpretability and data governance as ongoing challenges.

Keywords: Credit Risk, Natural Language Processing (NLP), BERT, RoBERTa, LLaMA, Transformer Models, Sentiment Analysis, Classification

1 Introduction

The primary focus of financial institutions is to assess and manage credit risk, necessitating the development and enhancement of complex credit risk models [1]. This literature review provides a comprehensive overview of how transformer-based NLP and Large Language Models are used for credit risk classification, highlighting key studies and developments in the field. Accurate credit risk classification is critical to ensure the stability and profitability of financial institutions within the broader scope of financial risk management [2]. Traditionally, credit scoring systems have relied largely on structured data and statistical methods to evaluate a borrower's credibility [2]. However, due to the rapid increase in unstructured financial text data, such as loan descriptions, transaction records, and customer communications, there is a growing recognition of the importance of utilizing these data to improve credit risk evaluations [3].

Natural Language Processing (NLP) models, particularly transformer-based architectures such as BERT, RoBERTa, and LLaMA, have enabled financial institutions to derive actionable insights from unstructured textual information. These models have shown significant improvements in classification tasks by capturing syntactic and semantic nuances in textual data, thus increasing the predictive accuracy of traditional credit risk models [4]. For instance, attention-based LSTM networks have been used to analyze borrower behavior patterns in peer-to-peer (P2P) lending platforms, resulting in enhanced credit scoring performance [4].

Recent studies have extended the application of transformer-based architectures, originally developed for natural language processing, to healthcare specifically for Alzheimer's disease risk assessment from speech transcripts demonstrating their effectiveness in early disease detection [5]. By leveraging models such as BERT (Bidirectional Encoder Representations from Transformers), researchers have demonstrated the efficacy of NLP in extracting meaningful risk indicators from loan narratives [6]. The integration of temporal models, including Temporal Convolutional Networks (TCN) and DilateFormer, with NLP techniques further refines classification accuracy by capturing temporal dependencies alongside textual semantics [7].

Despite these enhancements, several challenges remain. The interpretability of deep learning models, concerns regarding privacy, and the integration of insights derived from NLP into regulatory compliant frameworks continue to impede the widespread adoption of these approaches [8]. Furthermore, the current literature lacks a cohesive methodology for evaluating and benchmarking NLP-based credit risk classification models, leading to a fragmented understanding of their comparative effectiveness [2].

These enhancements and ongoing challenges highlight the evolving role of NLP in credit risk modeling within financial analytics. Research emphasizes how large language models and transformer architectures can enhance traditional credit scoring systems. However, to fully utilize these benefits, we need to resolve issues related to model transparency, data governance, and standardized evaluation. Overcoming these challenges is critical for the effective and ethical use of NLP in credit risk classification in real-world finance [1], [9].

The primary objective of this literature review is to critically analyze how the integration of natural language processing (NLP) techniques has transformed credit

risk model classification. It aims to evaluate how these technological advancements enhance risk management practices, improve the accuracy of credit scoring systems, and facilitate more informed decision-making within financial institutions. Specifically, the review intends to investigate the application of transformer-based NLP models in credit risk classification, with a particular focus on transformer-based approaches in practical financial settings. Additionally, it seeks to assess recent literature that highlights the contribution of unstructured data in enhancing the predictive accuracy of credit scoring models. Finally, the review will address key limitations, ethical challenges, and interpretability concerns associated with the use of NLP in credit risk modeling, emphasizing the importance of responsible and transparent model deployment within a regulated industry.

This literature review is guided by several core research questions. Firstly, it investigates how NLP techniques particularly advanced transformer-based models such as BERT, RoBERTa, and LLaMA have been applied to improve the classification of credit risk. Secondly, it explores the main benefits and challenges associated with incorporating unstructured data, such as customer communications and market sentiment, into credit risk models through NLP methodologies. Finally, the review seeks to identify which NLP-based methods or hybrid models demonstrate the most promise in enhancing the performance of credit risk models, while also addressing what limitations persist in their application, particularly in terms of transparency, accuracy, and ethical considerations.

2 Literature Review

This section provides background information on modern deep learning-based NLP models, specifically BERT, RoBERTa, and LLaMA 3.2 [10], [11], [12], within the specialized domain of credit risk classification. To maintain thematic coherence, we narrowed our review to studies that directly applied NLP and LLMs to financial and credit risk contexts [3], [6]. Articles that extended into unrelated NLP classification tasks were excluded to sharpen the review's focus. We began by outlining essential NLP principles and elucidate how these sophisticated models effectively process text data, particularly in the context of financial analysis.

Credit risk classification is a critical task that involves evaluating the likelihood of default by individuals or entities based on their historical performance and contextual information. The complexity of this undertaking lies in the need to analyze multifaceted inputs, including both unstructured text data (such as credit histories, financial reports, and market sentiment) and structured numerical data (like credit scores and income levels) [2]. In this landscape, models like BERT, RoBERTa, and LLaMA 3.2 stand out for their ability to extract actionable insights from vast amounts of unstructured text, enabling more informed decision-making [10], [11], [12].

BERT, an acronym for Bidirectional Encoder Representations from Transformers, leverages bidirectional context through the attention mechanism inherent to the Transformer architecture. Its pretraining objectives, which encompass Masked Language Modeling and Next Sentence Prediction, are crucial for capturing nuanced linguistic

patterns essential for accurate credit risk assessment [10]. By analyzing relationships within text, BERT can identify subtle shifts in language that may indicate potential credit risks.

Building on BERT's successes, RoBERTa (Robustly Optimized BERT Pretraining Approach) optimizes the training methodology by omitting certain tasks, such as Next Sentence Prediction, and incorporating dynamic masking strategies. These enhancements boost predictive accuracy and strengthen the model's ability to handle complex and large-scale datasets, thereby providing a more robust solution for credit risk classification [11]. LLaMA 3.2 (Large Language Model Meta AI) is designed as a specialized model for deployment in low-resource environments. With a focus on efficiency, it employs optimized data selection and advanced tokenization techniques that enhance scalability for extensive NLP tasks while maintaining high accuracy levels. This capability is particularly important in credit risk analysis, where efficiently processing and interpreting large volumes of financial data is crucial [12]. Central to these transformer-based models is the attention mechanism, which enables them to assess the importance of various words within a given context effectively. While this feature significantly enhances performance, it also presents challenges in terms of memory resource demands, as these models can be computationally intensive. In conclusion, we will examine potential strategies and solutions aimed at alleviating these memory constraints, facilitating the effective use of these state-of-the-art NLP models in credit risk classification tasks.

2.1 Credit Risk Models

Credit risk models are used to estimate the potential financial loss that a credit institution, such as a bank or peer-to-peer lender, might incur if a borrower defaults on repaying a loan. The most important component of a credit risk model is the probability of default, which is usually estimated statistically by employing credit scoring models. [1] examines how banks can adapt to manage credit risk amid economic uncertainty caused by rising interest rates, inflation, and geopolitical tensions. [9] provides an in depth examination of how machine learning enhances credit risk assessment and the role of Explainable AI in addressing its limitations. It highlights the high performance of machine learning algorithms in predicting credit risk, while emphasizing their lack of interpretability. The review discusses Explainable AI techniques like LIME and SHAP, which improve transparency by explaining model decisions, thereby fostering trust and ensuring fairness in lending practices.

[21] introduces the GPDBN-CRA model as a tool to assist financial institutions in evaluating loan applications. This innovative model integrates a Dynamic Bayesian Network to classify credit risk based on customer data, along with Genetic Programming for hyperparameter tuning, which enhances overall performance. Genetic programming is then used to optimize the parameters, leading to improved prediction accuracy. [22] proposes a novel approach to credit scoring by integrating blockchain technology and prospect theory. Blockchain ensures a secure, transparent, and tamper-proof credit scoring system, reducing reliance on centralized credit bureaus and enhancing data security. This innovative approach offers a secure, efficient

Table 1 Summary of Articles, Contributions, and Identified Gaps

Articles	Relation Between Articles	Contributions	Gaps
[13]	Proposes a new end-to-end framework for clinical information extraction (CIE) from Chinese radiology reports, termed normalized clinical information extraction (NCIE).	Presents the NCIE method for structured data extraction from radiology reports.	Dataset mainly uses chest X-ray and CT reports, limiting applicability to other anatomical domains.
[14]	Introduces a lexicon-based prompting method for financial sentiment analysis, addressing weaknesses in current domain-adaptive techniques.	Proposes a dictionary-driven prompt integration in language models for better sentiment classification.	Accuracy limited to polarity (positive/negative), as dictionary does not cover neutral or nuanced sentiment.
[15]	Develops a text-based metric for ESG risk exposure using a Word2vec-derived lexicon from ESG disclosures and news.	Establishes a method to quantify ESG risks from earnings call transcripts using text analytics.	Prior work emphasizes ESG performance, with limited focus on risk assessment and quantification.
[16]	Demonstrates the application of NLP to extract mechanistic information from scientific literature for toxicology.	It offers a proof-of-concept that NLP supports information extraction for modern toxicology.	The NLP pipeline requires automation to unify biological events and link them to semantic ontology terms.
[17]	This study presents a BERT-IDCNN-CRF model for entity extraction in R&D project abstracts.	Introduces a BERT-IDCNN-CRF model for scientific entity recognition in R&D project abstracts	Limitations in the annotated datasets due to the complexity of scientific texts.
[18]	Presents a comparative analysis of various approaches for extracting UML class diagrams from natural language specifications.	Showcases the accuracy of recent systems involving NLP and machine learning techniques.	Existing methods for translating requirements to UML diagrams have shortcomings.
[19]	Evaluates machine learning (ML) algorithms in predicting suicide risk.	Examine ML outcomes for predicting suicide and provide a list of suicide risk factors.	ML requires sufficient data to train and validate early identification of risk factors.
[20]	This paper introduces BERTeley, a Python package built on BERTopic, specifically designed for topic modeling of scientific articles.	BERTeley enhances topic modeling on scientific articles by addressing challenges posed by structural words and shared research goals.	Dilution of potential learnings due to stringent formatting and common words.
[18]	Presents a comparative analysis of various approaches for extracting UML class diagrams from natural language specifications.	Showcases the accuracy of recent systems involving NLP and machine learning techniques.	Existing methods for translating requirements to UML diagrams have shortcomings.
[19]	Evaluates machine learning (ML) algorithms in predicting suicide risk.	Examine ML outcomes for predicting suicide and provide a list of suicide risk factors.	ML requires sufficient data to train and validate early identification of risk factors.

method for assessing credit risk, potentially improving lending outcomes and minimizing defaults. [23] examines different credit scoring models to help financial institutions select the best method. It compares traditional techniques like logistic regression with advanced machine learning models, evaluating their accuracy, interpretability, and efficiency. The study highlights the practical implications of these models for various credit risk scenarios and explores trends like big data and explainable AI to improve transparency and trust in credit scoring.

2.2 Natural Language Processing

Natural language processing is centered on enabling computers to understand and process human languages. These models are crafted to learn a language's structure, syntax, and linguistic patterns or specific tasks by training on extensive text datasets relevant to the target language and objective [24]. The training of such models typically consists of three main steps. First, the text is segmented into smaller, manageable units through a process known as tokenization. These tokens are then transformed into a numerical format referred to as word embeddings, allowing for a more nuanced representation that supports mathematical operations [10]. Finally, a language model is trained using this tokenized and embedded dataset [24].

2.2.1 Language Modeling

A language model is a statistical representation of the probabilities associated with sequences of words or tokens. Based on a corpus of known words, it assigns likelihoods to each possible subsequent sequence of words or tokens $(t_1, t_2, \ldots, t_n)$ in a particular language L. In essence, it acts as a predictive model that identifies the most probable next words in a given context [24].

$$P(t_1, t_2, \ldots, t_n) = p(t_1)p(t_2|t_1) \cdots p(t_n|t_1, t_2, \ldots, t_{n-1})$$

$$= \prod_{i=1}^{n} p(t_i \mid t_1, t_2, \ldots, t_{i-1})$$

Training a language model involves utilizing a large unlabeled dataset to establish a foundational understanding of language. Subsequently, these models can be fine-tuned for specific downstream tasks using labeled data. The classical language modeling objective focuses on learning a left-to-right context, predicting the conditional probability of all preceding tokens $t_1, t_2, \ldots, t_{i-1}$ for a given token t_i. This traditional approach, which only generates predictions based on previously encountered words, is known as auto-regressive or unidirectional. It is particularly well-suited for tasks related to language generation [24].

A prevalent learning objective in natural language processing is the masked language model (MLM) objective. In this method, certain words or tokens within a sentence are randomly masked, and the model's goal is to predict these omitted tokens. By examining the context surrounding the masked words, the model assigns probabilities to the most likely candidates for replacement. This language modeling technique enables the model to learn in both left-to-right and right-to-left directions, facilitating

what is known as bidirectional reasoning. This capability is particularly advantageous for sentence-level tasks, including text classification, named entity recognition, sentence analysis, and question answering [10].

2.2.2 Encoder-Decoder Architectures

The Encoder-Decoder architecture is a neural network design for sequence modeling, particularly in NLP tasks like sentiment analysis and text classification, where input lengths can vary while outputs remain fixed. This architecture transforms input sequences of arbitrary length into output sequences Pakdaman et al. [25]. Traditional models struggle with this variability, requiring fixed lengths for training. The Encoder-Decoder consists of two networks, an encoder that converts an input sequence into a fixed-length vector representation, and a decoder that generates the output sequence from this vector. Both networks are trained together to accurately encode and reconstruct the target sequence.

2.3 The Transformer Model

Several pre-trained Transformer-based models, including LLaMA 3.2 [12], RoBERTa [11], and BERT [10], have set new standards for performance in various NLP applications since their introduction. RoBERTa enhances BERT's architecture by employing improved training methods, while LLaMA 3.2 focuses on efficiency and scalability in low-resource settings. The success of these models is largely attributed to their effective use of the attention mechanism, which enables strong and context-aware language understanding.

2.3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a language representation model created by researchers at Google AI Language in 2018. BERT was developed using deep learning techniques in conjunction with various other approaches. The architecture illustrated in Fig. 1 of the BERT model features a multi-layer bidirectional Transformer, which mirrors the original implementation of the Transformer but exclusively employs the encoder component [26].

[27] highlight that BERT has had a transformative effect on NLP, showcasing its effectiveness in various tasks such as text classification, sentiment analysis, named entity recognition, and question answering. It consistently surpasses earlier models like GPT and ELMo. The review also addresses proprietary BERT-based models created for specific applications, highlighting BERT's significance in establishing new benchmarks in NLP and inspiring further advancements within the field. BERT is designed to pre-train deep bidirectional representations from unlabeled text by simultaneously taking into account both the left and right contexts across all layers. Each sequence consistently begins with a special classification token. The final hidden state corresponding to this token acts as a unified sequence representation for classification tasks. Pairs of sentences are merged into a single sequence [28]. The study by [29] explores the use of an Enhanced BERT model to enhance sentiment analysis of Twitter data. Utilizing the Kaggle SMILE dataset, the research tailors BERT to accommodate the

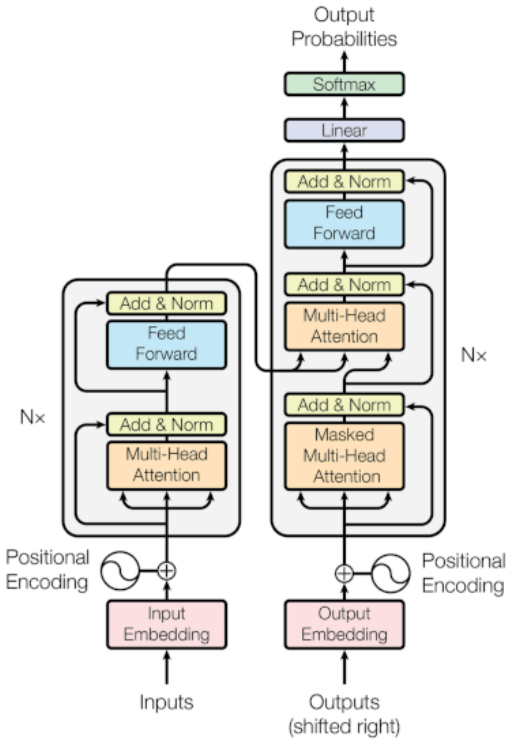

Fig. 1 Figure depicting the BERT model architecture. The figure is adapted from Rustam et al. [26], who developed a classification method for lecturer expertise areas using BERT.

informal nature of tweets, which frequently incorporate slang, acronyms, and emoticons. The Enhanced BERT model achieved an impressive accuracy rate of 96% in discerning sentiments such as happiness and sadness, highlighting its effectiveness. This research emphasizes the potential of Enhanced BERT in precisely analyzing the sentiments expressed in short, and informal texts.

2.3.2 RoBERTa

RoBERTa is an advanced natural language processing model created by Facebook AI in 2019, based on the BERT framework with several improvements aimed at enhancing its effectiveness [30]. The architecture of the RoBERTa model, as modified from [31], is depicted in Fig. 2.

2.3.3 LLAMA

The LLaMA $3.2-1B$ model, illustrated in Fig. 3, is a key component of the Large Language Model Meta AI (LLaMA) series, functioning as a robust tool for a wide range of natural language processing tasks. With 1 billion parameters, it effectively balances computational efficiency and performance, making it well-suited for both research and practical applications, especially in resource-constrained environments.

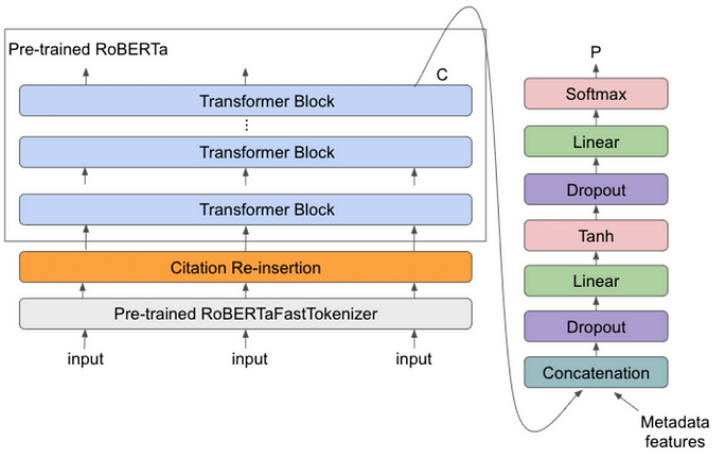

Fig. 2 Figure depicting the RoBERTa model architecture. The figure is adapted from Kumari [31], who described modifications made to BERT for NLP in the blog post RoBERTa: A Modified BERT Model for NLP.

Optimized for tasks such as text generation, summarization, and question answering, LLaMA $3.2 - 1B$ delivers impressive results while maintaining a lightweight profile compared to larger models in the series. Its design is tailored for users seeking an effective model that does not demand extensive computational power [32].

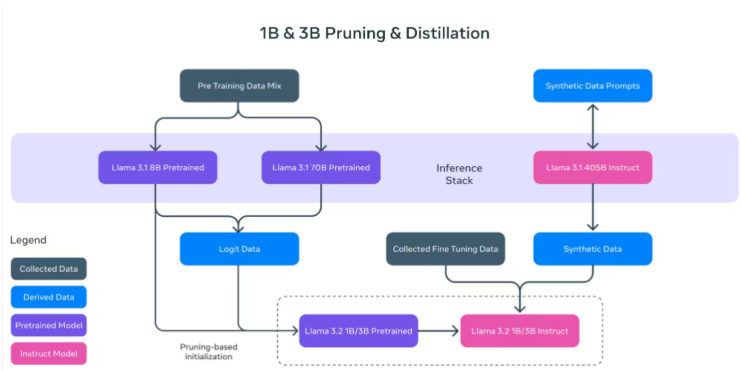

Fig. 3 Figure depicting the LLaMA 3.2 model architecture. The figure is adapted from the blog post, From Vision to Edge: Meta's LLaMA 3.2 Explained [33], which provides an overview of its advancements and applications.

[34] highlights the exceptional performance of LLaMA 3 in identifying subtle misinformation. It outperformed state-of-the-art competitors, achieving superior accuracy and precision because of its advanced contextual understanding and adaptability to complex narratives. Additionally, LLaMA 3 demonstrated remarkable efficiency in processing large datasets, rendering it suitable for real-time applications. Notably, the model demonstrated reasoning capabilities similar to human judgment, significantly

reducing both false positives and false negatives. These findings emphasize its potential as a powerful tool for combating misinformation. [35] discusses how large language models (LLMs) emulate human cognitive processes like reasoning and decision-making. It highlights advancements such as multi-layer attention mechanisms and transformer networks that enable effective handling of complex language tasks. The paper also addresses the importance of improving model interpretability and considers ethical issues related to bias, fairness, and misuse, while emphasizing LLMs' transformative potential in sectors like healthcare, finance, and education.

3 Comparative Analysis of Different Models and Datasets

Table 2 outlines the datasets employed in the reviewed studies. These datasets span from unstructured sources like financial tweets and news articles to structured stock market time series and legal judgments. Most are annotated with sentiment or emotion labels, while others use inferred labels like risk tolerance or psychometric traits. A few works rely on synthetic or proprietary data, limiting reproducibility and cross-study comparability. The table categorizes sources by type, language, size (where available), and annotation strategy, providing insight into the diversity and challenges of data collection in financial NLP. It also clarifies each dataset's intended analytical purpose, such as credit scoring, supplier evaluation, or emotion classification.

Table 3 presents a comparative overview of NLP models applied in financial and credit risk analysis. The models range from transformer-based architectures to convolutional and hybrid deep learning structures, each tailored to a specific problem such as sentiment analysis, credit scoring, or entity disambiguation. Most architectures integrate attention mechanisms or hybrid word embeddings to address polysemy, interpretability, and contextual representation in financial text. Some models, such as PsyCredit and PLWL, incorporate explainability techniques like LRP or dictionary filtering to enhance transparency. Performance across models varies, often based on task and dataset, but many demonstrate significant improvement over traditional baselines. The table also highlights which models emphasize explainability, accuracy, or application novelty.

Our comparative synthesis highlights how transformer-based models especially BERT and RoBERTa demonstrate superior performance across tasks such as sentiment classification and credit scoring. Hybrid approaches combining CNNs or GRUs with transformers further improve accuracy and adaptability, especially in imbalanced datasets. While LLaMA models show promise in low-resource contexts, BERT remains the most widely adopted. Notably, explainability is still lacking in many high performing models, with limited deployment of interpretability tools like SHAP or LIME in financial applications.

Table 2 Dataset Overview from Reviewed Articles on NLP for Credit and Financial Risk

Source	Data Type	Language	Annotation	Purpose
P2P Lending Descriptions [36]	Unstructured loan text	English	No manual annotation; LLM-derived risk indicators	Use borrower-written descriptions to construct interpretable credit risk indicators via Large Language Models (LLMs)
Social Media Posts [37]	User-generated text from P2P lending platforms	English	Self-supervised psychometric labeling	Infer borrowers' psychological traits to support credit risk evaluation
P2P Loan Applications [38]	User-generated loan descriptions	English	Labeled with default outcomes	Predict credit default using deep learning on textual applications
Local Government Work Reports [39]	Government documents	Chinese	Sentiment scores via NLP methods	Assess regional debt risk through sentiment analysis of official government reports
Bank Financial Data [40]	Structured financial indicators	Not specified	Labeled with default outcomes	Predict financial risk using structured data and deep learning models
Policy Documents on Mobile Payments [41]	Government policy documents	Chinese	Unsupervised entity and relationship extraction	Build knowledge graphs for mobile payment platforms to analyze risk and inform policy
Financial News Articles [42]	Financial news headlines	English	Labeled with sentiment polarity	Improve sentiment analysis accuracy using hybrid word embeddings and explainable AI techniques
Financial Text Data [14]	Financial documents	Not specified	Annotated with sentiment labels using lexicon-based prompts	Enhance financial sentiment analysis through lexicon-based prompt methods
Legal Judgments [43]	Legal case documents related to supply chain finance	Chinese	Annotated with credit risk outcomes	Utilize legal judgments to improve credit risk prediction in supply chain finance
Financial Twitter Data [44]	Tweets mentioning financial assets	English	Annotated with 'opportunity' and 'precaution' emotions per asset	Detect asset-specific financial emotions for market screening and decision-making
User-Generated Texts [45]	Text data from various sources	English	Labeled with financial risk tolerance levels	Profile financial risk tolerance using textual data
Financial Reports and Market Data [46]	Annual reports, financial indicators, and market data	English	Labeled with financial distress outcomes	Predict financial distress using an attentive and interpretable multimodal framework
Financial Transaction Graphs [47]	Transactional data represented as graphs	English	Labeled with fraudulent and non-fraudulent classes	Detect fraud in imbalanced datasets using graph neural networks

Table 3 Model Overview from Reviewed Articles on NLP for Credit and Financial Risk

Model	Architecture	Data Source	Input Features	Performance
LLM Risk Indicator [36]	Zero-shot LLM (GPT-3.5/BERT baseline)	P2P loan descriptions	Loan purpose and intention text authored by borrowers	Risk scores correlated with actual defaults; outperforms lexical baselines
PsyCredit [37]	Deep learning with LRP (Layer-wise Relevance Propagation)	Social media posts from P2P platforms	Linguistic psychometric signals	Outperformed baseline credit risk scoring models
Credit Default Predictor [38]	Deep Neural Network (DNN)	P2P loan applications	Descriptive features from borrower text	Achieved better accuracy than traditional financial models
LocalGov Debt Risk Model [39]	Deep learning NLP pipeline with sentiment analysis	Government work reports	Sentiment-based risk indicators	Effectively identified high-risk geographic regions
grcForest_XGB [40]	Ensemble of grcForest and XGBoost	Bank financial datasets	Structured financial indicators	Delivered superior predictive performance among benchmarked models
MobilePay Risk KG [41]	Deep learning for unsupervised knowledge graph construction	Mobile payment policy documents	Extracted entities and relationship triples from text	Enabled visual risk analysis and policy linkage via knowledge graphs
Hybrid Word Representation [42]	CNN with attention using hybrid word embeddings	Financial news headlines	Embeddings capturing semantic, syntactic, and polysemous cues	Achieved stronger sentiment classification than baseline approaches
CNN Risk Profiler [45]	CNN + psycholinguistic signals	Synthetic user-generated text	Linguistic markers + embeddings	Micro-F1 approx 0.51
BERTeley [20]	Transformer-based topic modeling with pre-trained scientific language models	Scientific article corpora	Sentence embeddings from scientific texts	Demonstrated superior coherence and topic quality
LLM4Jobs [48]	Unsupervised LLM-based framework	Job postings and resumes	Raw textual data	Outperformed existing unsupervised methods
LLM-based Chatbots for OSINT [49]	Various LLMs including ChatGPT, GPT4all, Dolly, Stanford Alpaca, Alpaca-LoRA, Falcon, and Vicuna	Twitter OSINT Dataset	Textual data from tweets	GPT-4 achieved an F1 score of 0.94 in binary classification; GPT4all achieved 0.90
Ensemble Oversampling with 3WD and SMOTE [50]	Bagging with logistic regression, three-way decisions, and SMOTE	Policy texts	Textual features for keyword extraction	Improved F-measure and G-mean

4 Methodology

This study uses a Systematic Literature Review (SLR) methodology to systematically identify, evaluate, and synthesize academic research on the integration of Natural Language Processing (NLP) models, specifically transformer-based architectures such as BERT, RoBERTa, and LLaMA, in credit risk classification tasks. The SLR uses Python based tools to ensure methodological transparency and reproducibility by integrating structured database querying with automated processes for filtering, cleaning, and scoring semantic similarity. To ensure comprehensive coverage, scholarly articles were sourced from three prominent databases: Scopus, ScienceDirect, and Web of Science. A standardized Boolean search string was crafted to explore these sources, focusing on papers on credit risk, financial risk, or loan default classification utilizing NLP or text mining techniques. The search was confined to English language publications released between 2015 and 2025. The query structure for Scopus is as follows:

```
TITLE-ABS-KEY (
    (credit risk
     OR financial risk
     OR loan default
     OR credit scoring)
    AND
    (natural language processing
     OR NLP
     OR text mining)
    AND
    (classification
     OR prediction)
)
AND PUBYEAR > 2015
AND PUBYEAR < 2025
AND (
    EXCLUDE(DOCTYPE, cp)
    OR EXCLUDE(DOCTYPE, re)
    OR EXCLUDE(DOCTYPE, cr)
    OR EXCLUDE(DOCTYPE, sh)
    OR EXCLUDE(DOCTYPE, ed)
)
AND (
    EXCLUDE(LANGUAGE, Chinese)
)
```

ScienceDirect's query was formulated using the basic search bar and fine-tuned for optimal keyword matching. Although the interface lacks the Boolean complexity

found in platforms like Scopus, it allows for a direct logical structure that effectively yields relevant peer-reviewed results. The search targeted the comprehensive full-text repository of Elsevier publications to locate articles that discuss NLP in the context of credit risk. The logical structure emphasized classification and prediction, specifically incorporating credit-related terminology. The search was also limited to publications in the English language that were released between 2015 and 2025. The following query demonstrates the syntax compatible with ScienceDirect's search functionality:

```
(
    credit risk
    OR financial risk
    OR loan default
    OR credit scoring
)
AND (
    natural language processing
    OR NLP
    OR text mining
)
AND (
    classification
    OR prediction
)
AND publicationYear:[2015 TO 2025]
AND language:English
```

The Web of Science platform offers a Topic Search (TS) feature for thorough querying of article titles, abstracts, author keywords, and Keywords Plus. The final query structure has been refined for universal compatibility and is implemented through the core collection interface. To narrow down the results, only publications from 2015 to 2025 in English were considered, and document types such as reviews and editorial materials were excluded. Below is the corresponding query that adheres to Web of Science's default logic syntax:

```
TS = (
    credit risk
    OR financial risk
    OR loan default
    OR credit scoring
)
AND TS = (
    natural language processing
    OR NLP
    OR text mining
)
AND TS = (
    classification
    OR prediction
)
AND PY = (2015-2025)
AND LA = (English)
NOT DT = (
    Meeting Abstract
    OR Editorial Material
    OR Review
)
```

4.1 Inclusion and Exclusion Criteria

To ensure thematic and methodological consistency, explicit eligibility criteria were applied:

Inclusion criteria:

- Application of transformer-based NLP models (BERT, RoBERTa, LLaMA, etc.) to credit risk or financial risk prediction
- Published between 2015 and 2025 in peer-reviewed journals or conferences
- Written in English

Exclusion criteria:

- Editorials, reviews, position papers, or non-empirical studies
- Studies not involving NLP/textual data

4.2 Screening, Deduplication, and Semantic Filtering

The initial search produced 284 unique records across the databases, which were then exported in BibTeX format and consolidated into a single bibliography. Following this export, a semantic filtering process was implemented to evaluate the relevance of each article against a predefined research theme. Of the total records, 217 articles were

excluded due to low semantic similarity, while 63 articles met the inclusion criteria. A comparative bar chart was created to visualize the proportions of included versus excluded articles, providing a clear representation of the data refinement process that underlies the methodological rigor of the review (Fig. 4).

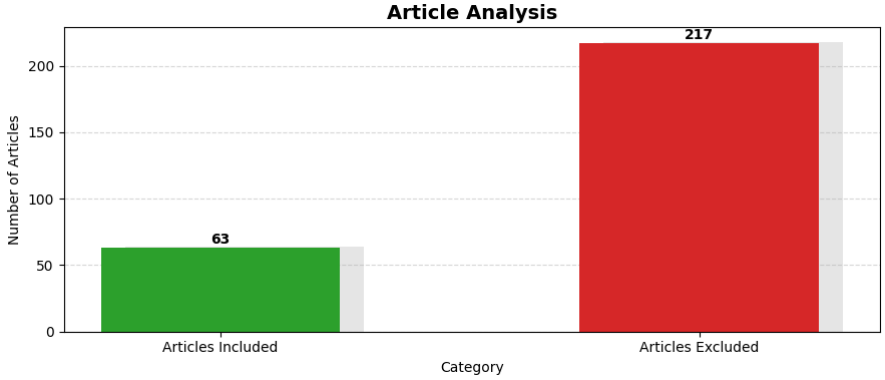

Fig. 4 Bar chart comparing the number of included versus excluded articles after semantic filtering.

A Python script utilizing the `bibtexparserlibrary` was used to load and parse the entries, while duplicates were identified and eliminated by string normalization of titles and unique identifiers.

```
import bibtexparser
from bibtexparser.bparser
import BibTexParser
from bibtexparser.customization
import homogenize_latex_encoding

def normalize(text):
    return text.strip().lower().replace('\n', '')
    if text else ''

def remove_duplicates(entries):
    seen_titles = set()
    unique_entries = []
    for entry in entries:
        title = normalize(entry.get('title', ''))
        if title not in seen_titles:
            seen_titles.add(title)
            unique_entries.append(entry)
    return unique_entries
```

After deduplication, 280 unique records remained. To refine the corpus to topically relevant studies, a semantic filtering step was introduced. Using the `SentenceTransformer` model `all-MiniLM-L6-v2`, Cosine similarity was calculated between the embeddings of each article derived from its title and keywords, and a target embedding that represents the central theme of the review. To reduce noise and computational costs during the initial filtering, abstracts despite containing important context were excluded. Future refinements may consider a tri-modal approach that includes abstracts for enhanced coverage.

```
from sentence_transformers
import SentenceTransformer, util
import torch

model = SentenceTransformer('all-MiniLM-L6-v2')

target_topic = (
 Optimising credit risk model classification using NLP,
    BERT, RoBERTa, and LLaMA
)
target_embedding = model.encode(
    target_topic, convert_to_tensor=True
)

SIMILARITY_THRESHOLD = 0.35
similar_articles = []

for entry in unique_entries:
    combined_text = "".join([
        entry.get(title, ""),
        entry.get(keywords, "")
    ])
    if not combined_text.strip():
        continue

    entry_embedding = model.encode(
        combined_text, convert_to_tensor=True
    )
    similarity = util.cos_sim(
        target_embedding, entry_embedding
    ).item()

    if similarity >= SIMILARITY_THRESHOLD:
        entry["similarity"] = round(similarity, 3)
        similar_articles.append(entry)
```

The threshold of 0.35 was established on the basis of preliminary experiments to ensure the inclusion of thematically relevant but semantically diverse articles. A lower threshold led to inclusion of unrelated works, while a higher one significantly reduced topical diversity. Consequently, only articles that met or surpassed a semantic similarity score of 0.35 were selected for detailed analysis, resulting in a curated collection of 63 studies. This rigorous filtering process ensured that the literature reviewed closely aligned with the thematic focus on credit risk modeling using NLP. In the interest of methodological transparency, the screening process adhered to PRISMA principles.

The key phases encompassed article identification, deduplication, semantic evaluation, and final inclusion based on relevance. Each stage contributed to refining the dataset into a robust evidence base. The distribution and flow of articles through these phases are visually represented in the Sankey-style PRISMA diagram provided below (Fig. 5).

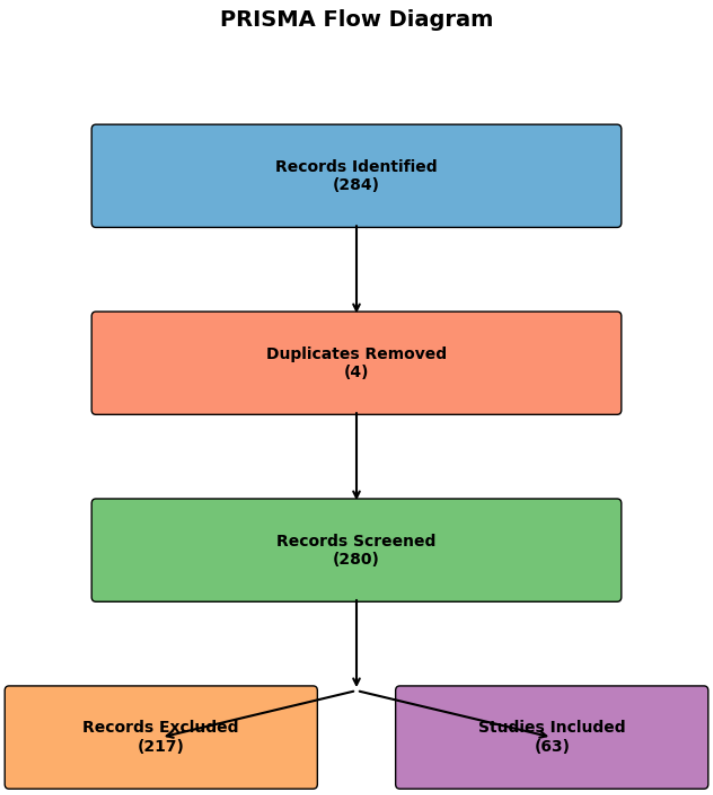

Fig. 5 PRISMA flow diagram outlining the systematic review process, including identification, screening, eligibility assessment, and inclusion of studies.

Figure 6, which illustrates publication trends, reveals a clear upward trajectory in relevant research output, particularly beginning in 2020. This trend highlights an increasing scholarly emphasis on the application of Natural Language Processing (NLP) within financial risk modeling. As depicted in Figure 6, the peak anticipated in 2025 signals a significant recent surge in academic engagement, likely driven by advancements in transformer-based models and the growing accessibility of financial text data. This pattern also reflects the broader integration of machine learning in finance, coinciding with various regulatory and technological changes. Overall, the

temporal analysis underscores the relevance and importance of this study within the dynamic research landscape.

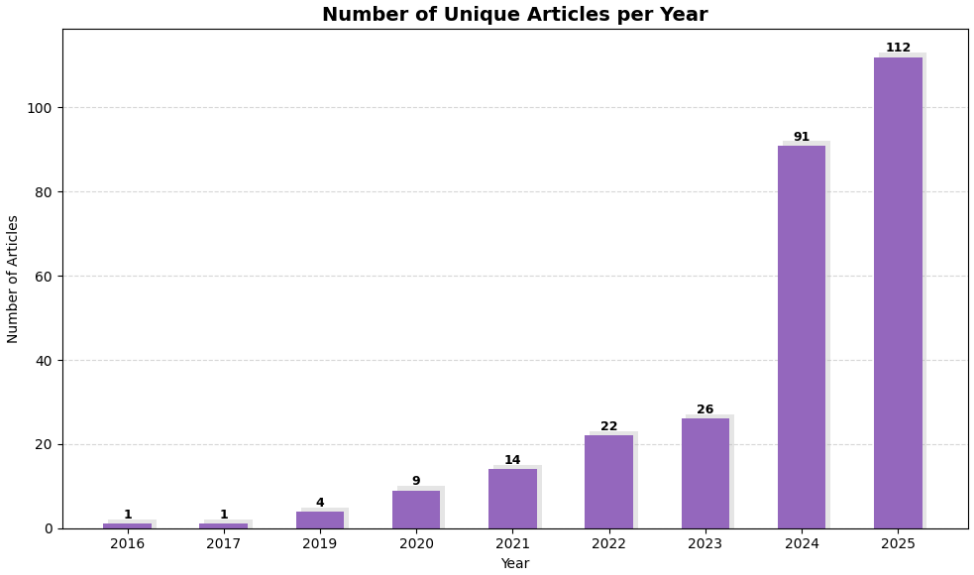

Fig. 6 Distribution and publication trends of selected studies over the years, illustrating the increasing academic interest in credit risk classification using NLP techniques.

The thematic structure of the reviewed literature is visually summarized in a word cloud (Fig. 7), which was generated by aggregating text from article titles and keywords. This visualization technique enabled the identification of key thematic patterns within the filtered corpus. Notable terms such as "LLM," "classification," "risk," , "based", "language model", and "prediction" appeared with high frequency, underscoring the central focus areas of the included studies. The prominence of these keywords highlights the importance of large language models and predictive analytics in the context of credit risk modeling. This thematic consistency further validates the article selection process and reinforces the alignment of the literature with the study's research objectives.

The integration of programmatic processing with scholarly review processes enhances methodological rigor and transparency. This hybrid approach utilizing NLP models, semantic embeddings, and structured screening frameworks ensures the inclusion of only the most relevant and high-quality studies. The resulting corpus serves as a solid foundation for analyzing current trends, assessing model efficacy, and understanding the evolving landscape of NLP in credit risk modeling.

4.3 Quality Assessment

Each included study was assessed for methodological quality and risk of bias using an adapted GRADE checklist. Criteria included:

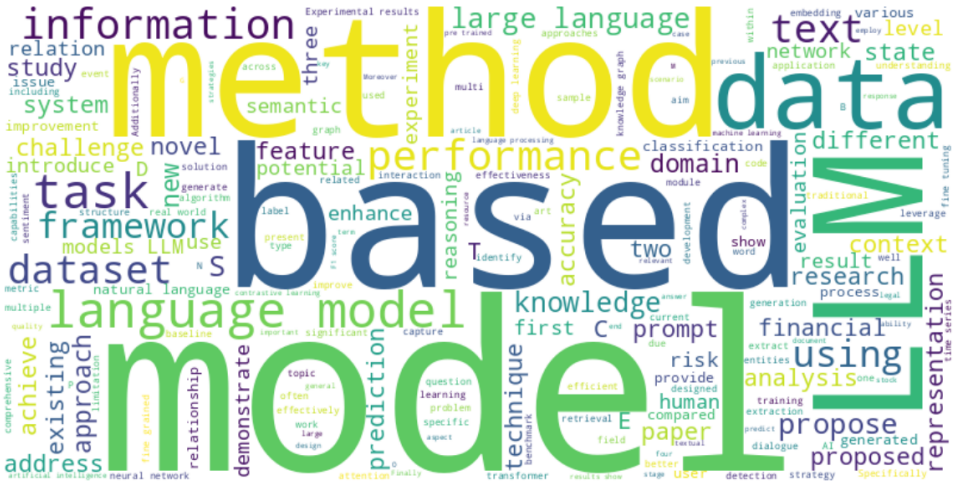

Fig. 7 Word cloud illustrating the most frequently occurring terms in the titles, abstracts, and keywords of the screened literature, highlighting key themes such as LLM, methods, models, classification, prediction, and risk.

- Dataset size and openness
- Clarity and reproducibility of methods
- Inclusion of baseline model comparisons
- Reporting of evaluation metrics

4.4 Data Extraction and Synthesis

For each of the 63 included articles, we extracted:

- Dataset type, size, and language
- Model architectures and training strategies
- Evaluation metrics and reported performance
- Use of explainability or fairness techniques
- Reported limitations or ethical considerations

5 Data Analysis and Discussion

5.1 Data Analysis

The paper [51] TCN-DilateFormer model shows high precision, recall, F1 score, and specificity, outperforming classical and state-of-the-art models in credit risk analysis. Parameter optimization revealed [1,2,4,8] as the optimal dilation factor combination and the original setting of [1,2,4] as the best scale configuration. The model exhibits tolerance to learning rate changes and low sensitivity to Gaussian noise variations, demonstrating robustness. Statistical tests confirm TCN-DilateFormer's superior performance compared to other models. [52] utilized multiple data sources, including the Airudi dataset, web scraping, and the RecSys Challenge 2017 dataset. The Airudi dataset contains job descriptions and resumes, while web scraping provided real-world

postings and candidate profiles. The authors used deep contextualized word embeddings and language models to focus on key feature extraction, such as skills and experience. A baseline model with transformer-based language models was developed to match candidates to job descriptions by assessing feature similarities. Performance was evaluated using metrics like NDCG and MRR, improving accuracy and explainability. The system ensured fair and unbiased matching, with traceability features to monitor decision-making and address discrimination concerns.

The study [26] introduced a novel method for classifying the areas of expertise of lecturers by analyzing the titles and abstracts of their scientific publications. By leveraging the Bidirectional Encoder Representations from Transformers BERT model, the researchers were able to convert unstructured text data into meaningful vector representations. This advanced deep learning model effectively classified the publications into 24 distinct scientific categories, achieving an impressive accuracy of 95.03% on the training data and 92.88% on the test data. This approach offers a more objective and data-driven framework for mapping lecturers' expertise, thereby enhancing the precision of academic talent management systems. [5] evaluated models using picture description test transcripts and the BERTLarge model with logistic regression achieved 88.08% classification accuracy. This improves the state-of-the-art by 2.48%. The sentence-level BERTLarge embedding method achieved the highest accuracy score 88.08%. The models based on the BERT family of embedders performed better than others. The best model has a Pearson correlation of 0.78 and a Spearman's rank correlation of 0.81 for the training phase. According to [4] the CNN-Transformer-WT model achieved a prediction accuracy of 0.969 on the Kaggle Dataset, 0.953, also surpassing other models. The risk recall for the CNN-Transformer-WT model was 0.947, higher than the GRU and BERT models. The model's comprehensive F1 score reached 0.955, exceeding other models' scores. The results confirm the effectiveness and superiority of the CNN-Transformer-WT model in financial risk prediction tasks.

The research [3] showed that LSTM-based models outperform traditional methods in credit risk assessment using P2P lending user data. The bidirectional LSTM (BLSTM) performs better than standard LSTM, as it captures sequence information from different directions. The attention-based LSTM (AM-LSTM) excels compared to standard BLSTM and BLSTM-Meanpool. The AM-LSTM model improved the KS value by 10.3% and AUC value by 3.08%. Models incorporating operation behavior data outperformed those without it, improving the KS value by 10.7% and the AUC value by 5.65%. This shows the advantages of using operation behavioral data in credit scoring. [53] suggests that NLP models can predict triage scores, need for admission, and critical illness with high accuracy, and can map free-text complaints to structured data. Multimodal models, incorporating both structured and free-text data, generally outperform models using only structured data. Free-text data alone can predict admission with high accuracy. Some studies show NLP models are more accurate than nurses in assigning triage scores. A deployed NLP model improved structured data capture. However, most studies have a high risk of bias, and few have been implemented in clinical practice.

In their study, [54] developed an algorithm to improve credit risk assessment, focusing on imbalanced datasets where loan defaults are a concern. They utilized

a dataset of 30,000 instances and used resampling techniques like SMOTE, SVM SMOTE, Random Undersampling, and ALL-KNN to enhance predictive accuracy. Various machine learning classifiers were used, including Deep Neural Networks, Support Vector Machines, K-Nearest Neighbors, and Artificial Neural Networks. The SVM model with ALL-KNN sampling achieved an impressive accuracy of 98.6% and a cross-entropy loss of 0.028. While emphasizing predictive accuracy, interpretable models like KNN and ANN also offered insights into the decision-making process, addressing concerns about fairness and explainability in credit risk assessments. This approach demonstrates the effectiveness of combining deep learning with resampling techniques for credit risk classification in imbalanced scenarios.

The paper [55] introduced a novel approach that combines the strengths of the Transformer-based RoBERTa model and Gated Recurrent Units (GRU) to improve sentiment analysis. The RoBERTa component effectively captures contextual word representations through its attention mechanism, while the GRU component adeptly models long-range dependencies in the text. To address challenges posed by imbalanced datasets, the authors employed data augmentation techniques, specifically oversampling minority classes using word embeddings, thereby enhancing the model's robustness and accuracy. Evaluations carried out on three widely used sentiment analysis datasets -IMDb, Sentiment140, and Twitter US Airline Sentiment - demonstrated the effectiveness of the model, achieving accuracies of 94.63%, 89.59% and 91.52%, respectively. These results underscore the potential of the RoBERTa-GRU hybrid model in advancing sentiment analysis tasks. [30] used the ELSTM-VC model, which combines an Extra Trees Classifier (ETC) with a Long Short-Term Memory (LSTM) network to detect explicit content in English song lyrics. They trained a dataset of 100 songs from Spotify, and the model achieved a 96% accuracy rate, outperforming existing machine learning models and encoding-decoding approaches. This automated method offers a more efficient and accurate solution for identifying inappropriate music content, protecting young listeners from exposure to explicit songs.

The authors of Alrowais et al. [56] developed a model that integrates the RoBERTa transformer with GloVe word embeddings to detect cyberbullying in tweets. This model was trained on a publicly available dataset focused on cyberbullying and achieved an impressive accuracy rate of 95%. It outperformed existing machine learning, deep learning, and transformer-based approaches that utilized FastText word embeddings. This research highlights the efficacy of combining transformer models with word embeddings to improve the detection of cyberbullying on social media platforms. The paper [29] investigated the application of an improved BERT model for sentiment analysis of tweets. They made use of the Kaggle SMILE dataset, which contains tweets labeled with various emotions, including happiness and sadness. Their enhanced BERT model achieved an impressive accuracy of 96% in classifying these sentiments, showcasing its effectiveness in comprehending the nuanced emotional content present in tweets.

Roumeliotis et al. [57] investigated how effective LLMs are at evaluating sentiment in product reviews from e-commerce sites. They assessed the performance of the GPT-3.5 and LLaMA-2 models, both in their pre-trained and fine-tuned versions, using a dataset of product reviews sourced from different e-commerce platforms. Their results

showed that fine-tuning greatly improved the accuracy of the models in sentiment classification, with GPT-3.5 achieving an accuracy rate of 92% and LLaMA-2 obtaining 89%. This study highlights the potential for LLMs to automate sentiment analysis in e-commerce, which can help businesses gain a better understanding of customer feedback. The paper by Jeyakarthic and Ramesh [21] presented the GPDBN-CRA model, which combines genetic programming with dynamic Bayesian networks to improve the assessment of credit risk. This model aids financial institutions in making decisions on loan applications by standardizing customer data and effectively evaluating creditworthiness. The GPDBN-CRA model exhibited better performance than conventional credit risk assessment approaches, providing a more precise and efficient tool for financial decision-making.

According to [58] the Monitoring and Early Warning Model Based on LSTM, Transformer, and Deep Learning model achieved a high accuracy across different datasets. 94.37% on the Fama-French Three-Factor Dataset and 93.84% on the CRSP Dataset. There are also significant improvements on the Compustat dataset with 95.13% accuracy and the World Bank dataset produced a 94.34% accuracy. The model's efficiency is evident through its low parameter count, short inference time of 213.58ms, and training time of 147.24 seconds on the Fama-French Three-Factor Dataset. Combining LSTM and Transformer boosts model accuracy and recall. The model can also detect risk signals earlier, enabling proactive risk management.

5.2 Discussion

The results of this systematic review indicate that the incorporation of Natural Language Processing (NLP) and Large Language Models (LLMs) into credit risk modeling signifies a noteworthy advancement in the way financial institutions analyze and interpret extensive amounts of unstructured data. Traditional models, which typically rely on structured inputs such as credit history, income level, and repayment behavior, are increasingly being enhanced and in some cases even surpassed by models that utilize textual sources like loan applications, social media posts, financial disclosures, and news articles [3], [4], [6], [37], [42], [44].

Among the most impactful models in natural language processing are BERT, RoBERTa, and LLaMA, which utilize attention mechanisms and pretraining strategies to draw out contextual and semantic insights from text [10], [12], [11]. These models excel in classification tasks related to credit scoring, default prediction, and fraud detection. For instance, BERT's bidirectional training facilitates a deeper comprehension of linguistic context, making it particularly effective at discerning nuanced sentiments within borrower narratives. RoBERTa enhances this capability with optimized pretraining and dynamic masking, while LLaMA provides scalability and efficiency suited for real-time risk assessment in resource-constrained environments [12].

In addition to these core models, the review highlights a trend toward hybrid architectures and multimodal frameworks that integrate NLP techniques with structured numerical features. Models like RoBERTa-GRU and CNN-Transformer hybrids demonstrate enhanced performance in sentiment analysis and risk classification, particularly when applied to intricate datasets that encompass both text and time-series

data [4]. Attention-based models, such as ACWGAN-GPSA and the RCMA fraud detection system, exemplify how detailed modeling of financial documents and legal texts can reveal concealed risk signals [43], [47].

Despite these advancements, several challenges continue to exist. Ethical and regulatory considerations are vital for real world deployment of NLP-based credit risk models. Fairness in lending is a primary concern, particularly given evidence of algorithmic bias against certain demographic groups. Model interpretability remains a significant concern, particularly in regulated sectors such as finance, where transparency is crucial. Explainable AI (XAI) techniques, including LIME and SHAP, are increasingly being utilized to tackle this issue; however, they still fall short when it comes to capturing the decision-making logic of highly complex models [9]. Additionally, problems related to data imbalance, especially the underrepresentation of default cases, highlight the need for techniques like SMOTE, ensemble oversampling, and adaptive weighting to achieve fair and accurate classifications [23], [50].

Ethical considerations are another significant theme, with many studies underscoring the potential for algorithmic bias, privacy violations, and misuse of sensitive financial or personal data [9]. This necessitates the development of standardized evaluation protocols and ethical guidelines for deploying LLMs in risk-sensitive domains [35].

6 Conclusion

This review offers a thorough exploration of how Natural Language Processing and Large Language Models are transforming the field of credit risk classification. By facilitating the integration of unstructured textual data and enhancing the semantic interpretation of financial narratives, models such as BERT, RoBERTa, and LLaMA represent a significant shift in risk modeling practices [10], [11], [12]. These advancements not only enhance predictive accuracy but also introduce new avenues for analysis, including psychometric profiling, sentiment trajectory modeling, and multimodal fraud detection [4], [37], [45], [46].

The integration of Natural Language Processing (NLP) into credit risk modeling presents several challenges. Key concerns regarding interpretability, fairness, regulatory compliance, and scalability continue to dominate discussions in both academic and industrial circles [8], [1], [9]. There is an evident necessity for more transparent models that can earn the trust of regulators and end-users alike. Additionally, it is crucial to address the prevention of existing biases and to safeguard user privacy within these models [22], [47]. The incorporation of Natural Language Processing (NLP) into credit risk modeling brings forth a number of challenges. Prominent issues related to interpretability, fairness, regulatory compliance, and scalability remain central to discussions in both academic and industry settings. There is a clear need for more transparent models that can foster trust among regulators and end-users alike. Moreover, it is essential to tackle the prevention of existing biases and ensure equitable outcomes.

In summary, although NLP and LLMs present significant opportunities for enhancing credit risk assessment, fully harnessing their advantages necessitates a balanced

strategy that combines technical innovation with ethical responsibility [2], [9], [35]. The future of credit risk modeling rests on the responsible and transparent application of these transformative technologies.

Funding

The authors received no financial support for the research.

References

[1] Pathak, K., Rougeaux, C., Singh, H.: Navigating Economic Uncertainty: New Guidance for Credit Risk Management (2023). https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/navigating-economic-uncertainty-new-guidance-for-credit-risk-management

[2] Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.: Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research **(doi:10.1016/j.ejor.2015.05.030)** (2015) https://doi.org/10.1016/j.ejor.2015.05.030

[3] Wang, C., Han, D., Liu, Q., Luo, S.: A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism lstm. IEEE ACCESS **7**, 2161–2168 (2019) https://doi.org/10.1109/ACCESS.2018.2887138

[4] Jin, J., Zhang, Y.: Innovation in financial enterprise risk prediction model: A hybrid deep learning technique based on cnn-transformer-wt. JOURNAL OF ORGANIZATIONAL AND END USER COMPUTING **36**(1) (2024) https://doi.org/10.4018/JOEUC.361650

[5] Roshanzamir, A., Aghajan, H., Soleymani Baghshah, M.: Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech. BMC MEDICAL INFORMATICS AND DECISION MAKING **21**(1) (2021) https://doi.org/10.1186/s12911-021-01456-3

[6] Sanz-Guerrero, M., Arroyo, J.: Credit risk meets large language models: Building a risk indicator from loan descriptions in p2p lending. Inteligencia Artificial **28**(75), 220–248 (2025) https://doi.org/10.4114/intartif.vol28iss75pp220-248 . Cited by: 0; All Open Access, Gold Open Access

[7] Shen, C., Wu, J.: Research on credit risk of listed companies: a hybrid model based on tcn and dilateformer. SCIENTIFIC REPORTS **15**(1) (2025) https://doi.org/10.1038/s41598-025-86371-7

[8] Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., Vasilakos, A.V.: Privacy and security issues in deep learning: A survey. IEEE ACCESS **9**, 4566–4593 (2021) https://doi.org/10.1109/ACCESS.2020.3045078

[9] Heng, Y., Subramanian, P.: A Systematic Review of Machine Learning and Explainable Artificial Intelligence (XAI) in Credit Risk Modelling, pp. 596–614 (2022). https://doi.org/10.1007/978-3-031-18461-1_39

[10] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

[11] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv **abs/1907.11692** (2019)

[12] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv (2023) https://doi.org/10.48550/arXiv.2302.13971

[13] Xu, Q., Xu, X., Zhou, C., Liu, Z., Huang, F., Li, S., Zhu, L., Bai, Z., Xu, Y., Hu, W.: Towards normalized clinical information extraction in chinese radiology report with large language models. Expert Systems with Applications **271**, 126585 (2025) https://doi.org/10.1016/j.eswa.2025.126585

[14] Lin, W., Liao, L.-C.: Lexicon-based prompt for financial dimensional sentiment analysis. Expert Systems with Applications **244**, 122936 (2024) https://doi.org/10.1016/j.eswa.2023.122936

[15] Du, F., Bao, C., Meng, Q., Hui, Y.: Text-based measure of esg risk exposure. Procedia Computer Science **242**, 693–700 (2024) https://doi.org/10.1016/j.procs.2024.08.142 . 11th International Conference on Information Technology and Quantitative Management (ITQM 2024)

[16] Corradi, M.P.F., de Haan, A.M., Staumont, B., Piersma, A.H., Geris, L., Pieters, R.H.H., Krul, C.A.M., Teunis, M.A.T.: Natural language processing in toxicology: Delineating adverse outcome pathways and guiding the application of new approach methodologies. Biomaterials and Biosystems **7**, 100061 (2022) https://doi.org/10.1016/j.bbiosy.2022.100061

[17] Hu, Y., Yu, J., Wang, K., Liao, F.: Entity extraction for research and development project knowledge graph. Procedia Computer Science **242**, 17–24 (2024) https://doi.org/10.1016/j.procs.2024.08.221 . 11th International Conference on Information Technology and Quantitative Management (ITQM 2024)

[18] Babaalla, Z., Abdelmalek, H., Jakimi, A., Oualla, M.: Extraction of uml class diagrams using deep learning: Comparative study and critical analysis. Procedia Computer Science **236**, 452–459 (2024) https://doi.org/10.1016/j.procs.2024.05.053 . International Symposium on Green Technologies and Applications (ISGTA'2023)

[19] Xu, L., Li, Z.: A new appraisal model of second-hand housing prices in china's first-tier cities based on machine learning algorithms. COMPUTATIONAL ECONOMICS **57**(2), 617–637 (2021) https://doi.org/10.1007/s10614-020-09973-5

[20] Chagnon, E., Pandolfi, R., Donatelli, J., Ushizima, D.: Benchmarking topic models on scientific articles using berteley. Natural Language Processing Journal **6**, 100044 (2024) https://doi.org/10.1016/j.nlp.2023.100044

[21] Jeyakarthic, M., Ramesh, R.: Genetic programming with dynamic bayesian network based credit risk assessment model, pp. 845–850 (2023). https://doi.org/10.1109/ICAIS56108.2023.10073788

[22] Hassija, V., Bansal, G., Chamola, V., Kumar, N., Guizani, M.: Secure lending: Blockchain and prospect theory-based decentralized credit scoring model. IEEE Transactions on Network Science and Engineering **PP** (2020) https://doi.org/10.1109/TNSE.2020.2982488

[23] Jukna, D.: Credit scoring methods comparison - a review. (2022)

[24] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Amodei, D.: Language models are few-shot learners. ArXiv (2020) https://doi.org/10.48550/arXiv.2005.14165

[25] Pakdaman, Z., Koochari, A., Sharifi, A.: Content-aware sentiment understanding: cross-modal analysis with encoder-decoder architectures. Journal of Computational Social Science **8** (2025) https://doi.org/10.1007/s42001-025-00374-y

[26] Rustam, D., Suhendra, A., Harmanto, S., Suhatril, R., Saputra, D., Tafsili, R., Prasetya, R.: Development of classification method for lecturer area of expertise based on scientific publication using bert. International Journal on Advanced Science, Engineering and Information Technology **14**, 894–905 (2024) https://doi.org/10.18517/ijaseit.14.3.19893

[27] Koroteev, M.: Bert: A review of applications in natural language processing and understanding (2021) https://doi.org/10.48550/arXiv.2103.11943

[28] Chotirat, S., Meesad, P.: Natural Language Processing with "More Than Words – BERT", pp. 108–116 (2021). https://doi.org/10.1007/978-3-030-79757-7_11

[29] Mann, S., Arora, J., Bhatia, M., Sharma, R., Taragi, R.: Twitter Sentiment Analysis Using Enhanced BERT, pp. 263–271 (2023). https://doi.org/10.1007/978-981-19-6581-4_21

[30] Chen, X., Aljrees, T., Umer, M., Karamti, H., Tahir, S., Abuzinadah, N., Alnowaiser, K., Eshmawi, A., Mohamed, A., Ashraf, I.: A novel approach for

explicit song lyrics detection using machine and deep ensemble learning models. PeerJ Computer Science **9**, 1469 (2023) https://doi.org/10.7717/peerj-cs.1469

[31] Kumari, K.: RoBERTa: A Modified BERT Model for NLP (2023). https://www.comet.com/site/blog/roberta-a-modified-bert-model-for-nlp/

[32] Hugging Face: LLaMA 3.2 Model (1B) Release. Accessed: 2024-11-19 (2024). https://huggingface.co/meta-llama/Llama-3.2-1B

[33] Encord: From Vision to Edge: Meta's LLaMA 3.2 Explained. Accessed: 2024-11-19 (2024). https://encord.com/blog/lama-3-2-explained/

[34] Stefan, R., Brad, R.: Llama 3 vs. state-of-the-art large language models: Performance in detecting nuanced fake news. Computers **13**, 292 (2024) https://doi.org/10.3390/computers13110292

[35] Sarfati, R., Liu, T., Boullé, N., Earls, C.: Lines of thought in large language models. ArXiv (2024) https://doi.org/10.48550/arXiv.2410.01545

[36] Yuan, J., Wei, L.: The impact of sentiment consistency on bank risk: A dual perspective from bank managers and credit rating experts. Procedia Computer Science **242**, 789–796 (2024) https://doi.org/10.1016/j.procs.2024.08.196 . 11th International Conference on Information Technology and Quantitative Management (ITQM 2024)

[37] Yang, K., Yuan, H., Lau, R.Y.K.: Psycredit: An interpretable deep learning-based credit assessment approach facilitated by psychometric natural language processing. EXPERT SYSTEMS WITH APPLICATIONS **198** (2022) https://doi.org/10.1016/j.eswa.2022.116847

[38] Brahma, A., Goldberg, D.M., Zaman, N., Aloiso, M.: Automated mortgage origination delay detection from textual conversations. Decision Support Systems **140**, 113433 (2021) https://doi.org/10.1016/j.dss.2020.113433

[39] Guo, Y., Li, Y., Qian, Y.: Local government debt risk assessment: A deep learning-based perspective. INFORMATION PROCESSING & MANAGEMENT **59**(3) (2022) https://doi.org/10.1016/j.ipm.2022.102948

[40] Tao, F., Wang, W., Lu, R.: A deep learning-based sentiment flow analysis model for predicting financial risk of listed companies. Engineering Applications of Artificial Intelligence **150** (2025) https://doi.org/10.1016/j.engappai.2025.110522 . Cited by: 0

[41] Xia, H., Wang, Y., Gauthier, J., Zhang, J.Z.: Knowledge graph of mobile payment platforms based on deep learning: Risk analysis and policy implications. Expert Systems with Applications **208**, 118143 (2022) https://doi.org/10.1016/j.eswa.2022.118143

[42] Adhikari, S., Thapa, S., Naseem, U., Lu, H.Y., Bharathy, G., Prasad, M.: Explainable hybrid word representations for sentiment analysis of financial news. Neural Networks **164**, 115–123 (2023) https://doi.org/10.1016/j.neunet.2023.04.011

[43] Wang, W., Chen, Y., Wang, L., Xiong, Y.: Developing the value of legal judgments of supply chain finance for credit risk prediction through novel acwgan-gpsa approach. Transportation Research Part E: Logistics and Transportation Review **196** (2025) https://doi.org/10.1016/j.tre.2025.104020 . Cited by: 0

[44] García-Méndez, S., de Arriba-Pérez, F., Barros-Vila, A., González-Castaño, F.J.: Targeted aspect-based emotion analysis to detect opportunities and precaution in financial twitter messages. Expert Systems with Applications **218**, 119611 (2023) https://doi.org/10.1016/j.eswa.2023.119611

[45] Xing, F.: Financial risk tolerance profiling from text. INFORMATION PROCESSING & MANAGEMENT **61**(4) (2024) https://doi.org/10.1016/j.ipm.2024.103704

[46] Che, W., Wang, Z., Jiang, C., Abedin, M.Z.: Predicting financial distress using multimodal data: An attentive and regularized deep learning method. Information Processing & Management **61**(4), 103703 (2024) https://doi.org/10.1016/j.ipm.2024.103703

[47] Boyapati, M., Aygun, R.: Balancergnn: Balancer graph neural networks for imbalanced datasets: A case study on fraud detection. Neural Networks **182**, 106926 (2025) https://doi.org/10.1016/j.neunet.2024.106926

[48] Li, N., Kang, B., De Bie, T.: Llm4jobs: Unsupervised occupation extraction and standardization leveraging large language models. Knowledge-Based Systems **316**, 113302 (2025) https://doi.org/10.1016/j.knosys.2025.113302

[49] Shafee, S., Bessani, A., Ferreira, P.M.: Evaluation of llm-based chatbots for osint-based cyber threat awareness. Expert Systems with Applications **261**, 125509 (2025) https://doi.org/10.1016/j.eswa.2024.125509

[50] Liang, D., Yi, B., Cao, W., Zheng, Q.: Exploring ensemble oversampling method for imbalanced keyword extraction learning in policy text based on three-way decisions and smote. Expert Systems with Applications **188**, 116051 (2022) https://doi.org/10.1016/j.eswa.2021.116051

[51] Shen, C., Wu, J.: Research on credit risk of listed companies: a hybrid model based on tcn and dilateformer. Scientific Reports **15**(1) (2025) https://doi.org/10.1038/s41598-025-86371-7 . Cited by: 0; All Open Access, Gold Open Access

[52] Barrak, M., Adams, B., Zouaq, A.: Toward a traceable, explainable, and fairjd/resume recommendation system. ArXiv (2022) https://doi.org/10.48550/arXiv.2202.08960

[53] Shen, Q., Zhang, X., Ren, H., Guo, Q., Yi, Z.: Knowledge-embedded large language models for emergency triage. Knowledge-Based Systems, 113431 (2025) https://doi.org/10.1016/j.knosys.2025.113431

[54] Mahbobi, M., Kimiagari, S., Marriappan Vasudevan, M.V.: Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks. Annals of Operations Research **330**, 1–29 (2021) https://doi.org/10.1007/s10479-021-04114-z

[55] Tan, K., Lee, C.-P., Lim, K.: Roberta-gru: A hybrid deep learning model for enhanced sentiment analysis. Applied Sciences **13**, 3915 (2023) https://doi.org/10.3390/app13063915

[56] Alrowais, F., Jamjoom, A., Karamti, H., Umer, M., Alsubai, S., Kim, T.-H., Ashraf, I.: Robertanet: Enhanced roberta transformer based model for cyberbullying detection with glove features. IEEE Access **PP**, 1–1 (2024) https://doi.org/10.1109/ACCESS.2024.3386637

[57] Roumeliotis, K., Tselikas, N., Nasiopoulos, D.: Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. Natural Language Processing Journal **6**, 100056 (2024) https://doi.org/10.1016/j.nlp.2024.100056

[58] Song, Y., Du, H., Piao, T., Shi, H.: Research on financial risk intelligent monitoring and early warning model based on lstm, transformer, and deep learning. JOURNAL OF ORGANIZATIONAL AND END USER COMPUTING **36**(1) (2024) https://doi.org/10.4018/JOEUC.337607