

https://dspace.org/

Exploring Personality Structure in South Africa: A Text Mining Approach

Gama, Beauty

2024-03-15
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/42463
Gama, Beauty. (2024). Exploring Personality Structure in South Africa: A Text Mining Approach [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/42463

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

Exploring Personality Structure in South Africa: A Text Mining Approach

By

Beauty Gama

Student number

1491833

Supervised by

Supervisor: Professor Sumaya Laher

Co-supervisor: Professor Rod Alence

Submitted on:

15 March 2024

Research Report

Submitted in partial fulfilment of the requirements for

the Degree of Master of Arts in E-Science (SOSS7094A)

in the School of Social Sciences, Faculty

of Humanities, at the University of the Witwatersrand, Johannesburg

Declaration

I, Beauty Sibongile Lindiwe Gama declare that this research report is my own, unaided work. It is being submitted in partial fulfilment of the requirements for the Masters of Arts in E-science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.



Signature

15 March 2024

Date

Abstract

Physical expression, behavioural attributes and social relations of an individual can often be studied through personality traits. This has made personality research a relevant aspect of gaining a deeper understanding of people in various contexts, for clinical reasons as well as social relatability. Trait theory has been fundamental in utilizing statistical methods such as factor analysis to construct the personality models that currently exist. The Five Factor Model (FFM) is amongst the most widely accepted of these trait theory models. Personality assessment instruments are developed as operationalisations of these models. These include the Goldberg Adjective Checklist, the South African Personality Inventory (SAPI), and the Chinese Personality Assessment Inventory (CPAI). Recently, naturally occurring data like social media statuses or Facebook Posts are being considered as data examining personality structure. This study aims to explore personality structure data obtained from South African literary texts and text mining techniques. Various techniques of text mining such as parts of speech tagging, and unsupervised and supervised LDA topic modelling were applied to 60 South African literary texts. While topic modelling showed limitations when used in an unsupervised manner, when guided by thematic clusters it presented comprehensible trait classifications that fit with the clusters as defined by the FFM. The instances where there was no fit corresponded with the literature which demonstrates poor fit for those constructs in African constructs. The results also showed that there is a difference in the expression of personality traits between men and women with the differences concurring with those found in the broader literature on gender differences across personality. While the text corpus for this study was small, there is evidence to suggest that text mining techniques could be used to assist in research on personality structure. Text mining is an approach that requires further research as it can be useful in dealing with large data that is naturally occurring to provide a better contextual exploration of personality.

Keywords:

Personality; Text Mining; Supervised Topic Modelling; Unsupervised Topic Modelling; SAPI; Latent Dirichlet Allocation (LDA); Five Factor Model

Acknowledgements

I would like to express my deepest gratitude to the following people and organisations that supported me throughout the degree:

- My deepest thank you to the DST-funded National e-Science Postgraduate Teaching and Training Platform (NEPTTP) for financial and academical support as well as the opportunity to be part of the course.
- To my supervisors, Professor Sumaya Laher and Professor Rod Alence, for always believing in my interests and supporting me throughout even with ups and downs you have been the mentor that has allowed me to grow so much. Thank you so much.
- My family, Thozamile Gama, has given me the best support structure I could ever ask for. The daily motivations and reminders of what I am capable of, I am forever grateful for that. Sindi Mabaso and Lindani Mabaso for always encouraging and supporting me and forever being proud of me. And my best friend Bridget Kabuya, for showing your unwavering care and pride.
- Hello Ara Team, for amazing colleagues and a family that has taken care of me and supported me throughout. I am so grateful for the kindness you have all shown me.
- My father Thulani Gama, has instilled values that I continuously live by and the love for education is what has put me on this journey. You might not be around but I feel your protection, pride and guidance every day and your spirit lives on.
  Wena wekusa nelilanga.
- My overall gratitude to Simakadze and Emalangeni, for the protection, the guidance and for lending me a hand whenever I am on my knees. Thank you for the grace you have shown me, I am always out of words when it comes to thanking you.

Table of Contents

Declaration 2
Abstract 3
Keywords: 3
Acknowledgements 4
Chapter 1 – Background 9
1.1 Introduction 9
1.2 Rationale 10
1.3 Research Aims 12
1.4 Study Outline 12
Chapter 2 – Literature Review 13
2.1 Introduction 13
2.2 Personality 13
2.3 Personality traits and language 14
2.3.1 Researcher-based trait collection 14
2.3.2 Computer-based trait collection 15
2.4 Personality models 17
2.5 Personality Assessment 20
2.6 Application of personality theory and models: Recent research 22
2.7 The Chinese Personality Assessment Inventory (CPAI) 23
2.8 Development of the South African Personality Inventory (SAPI) 25
2.9 Text mining 27
2.9.1 Information retrieval 27
2.9.2 Text Parsing and Filtering 28
2.9.3 Text transformation 28
2.9.4 Text analysis 29
2.10 Topic Modelling versus Factor Analysis 30
2.11 Research Questions 30
2.12 Conclusion 31
Chapter 3 – Methods 32
3.1 Introduction 32
3.2 Research Design 32
3.3 Sample 32
3.4 Procedure 33
3.5 Ethical considerations 33
3.6 Data Analysis 34
3.6.1 Data Cleaning 34

3.6.2 Coding Process 35
3.6.3 Analyses 36
3.7 Conclusion 37
Chapter 4 – Results 38
4.1 Introduction 38
4.2 Descriptive textual analysis 38
4.2.1 Corpus 38
4.3 Personality trait distribution 43
4.3.1 Trait variation in different gender identities 43
4.4 Topic Modelling – General Corpus 48
4.4.1 Unsupervised Method – LDA 48
4.4.2 Semi-supervised Method 51
4.5 Topic Modelling – Adjective distribution 65
4.5.1 Unsupervised LDA 65
4.5.2 Supervised LDA 68
4.6 Conclusion 69
Chapter 5 – Discussion 70
5.1 Introduction 70
5.2 Research Question 1 – How are personality descriptors classified using text mining? 70
5.2.1 Corpus analysis 70
5.2.2 Textual readability 71
5.2.3 Part of speech tagging - Adjectives 71
5.2.4 Part of speech tagging – Gender differences 73
5.2.5 Topic modelling 74
5.3 Research Question 2 – Are there different personality clusters produced using topic modelling? 77
5.3.1 Topic Modelling – General Corpus 77
5.3.2 Topic Modelling – Adjective Distribution 78
5.4 Research Question 3 – What are the common themes found in the South African literary text in comparison to existing personality classifications? 79
5.4.1 Supervised Text Mining - Frequency 79
5.4.2 Supervised Text Mining – LDA 81
5.5 Limitations 84
5.5.1 Large Corpus Issue 84
5.5.2 Noisy Data 84
5.5.3 Optimal number of topics and interpretation 85
5.5.4 Exploration of Other Topic Models 85
5.6 Recommendations 85

5.6.1 Large Data Issue – Split Corpus 85
5.6.2 Noisy Data – Noise Controlling Approach 86
5.6.3 Exploration of Other Topic Models – Comparing different models 86
5.6.4 Future Research – Personality Construction, NLP, and Generative AI 86
5.7 Conclusion 87
References 89
Appendix – Corpus full details 108

Chapter 1 – Background

1.1 Introduction

The empirical study of personality traits can be encompassed in two categories of common-sense discourse on human nature (John & Robins, 2021; Matthews et al., 2009). These include using natural language as a trait descriptor for people and bringing awareness of the generalities of personality traits among individuals. Natural language terms can be employed to represent individual personality differences (De Raad & Mlacic, 2015; John & Robins, 2021; Matthews et al., 2009). Though McCrae et al. (2000) argue that personality traits cannot be observed, they can be inferred using valid linguistic trait indicators from existing patterns of behaviours. Using statistical measures similar to factor analysis allowed for the reduction of rating schemes into three factors captured by the Eysenck Personality Questionnaire (Eysenck, 1953; Matthews et al., 2009). As such, the development of personality scales like (16 NEO-PI-R and the Five-Factor model) has allowed for a scientific manner of personality observation through traits like extroversion and neuroticism as well as their comparability among individuals.

The use of lexical strategies as a means of studying implicit personality structures provides a way of better understanding the concept of personality. However, the cross-cultural relevance of such measures is often called into question. This is because they were developed using the English lexicon and validated using WEIRD (Western, Educated, Industrialised, Rich and Democratic) samples (Henrich et al., 2010; Morton et al., 2019). Further, studies on lexical strategies prove that they do not support the universality of personality traits (Cheung et al., 2008; Nel et al., 2012; Zeinoun et al., 2018). De Fruyt and Wille (2013) outline that personality assessment, especially in professional settings, will represent more globalised economies. They acknowledge that multicultural contexts present issues beyond individual personality differences but include contextual relatability. Questions such as "What kind of trait model should one use to describe an individual's personality within and across cultural contexts?" and "What norms should one use when comparing individuals from diverse cultural backgrounds?" (De Fruyt & Wille, 2013, p. 2) have come up in cross-cultural research.

Nel et al. (2012), in exploring the personality structures in eleven languages of South Africa, argue that the Five-Factor model may not be complete since it fails to consider social elements of personality. This can also be observed in a study done in the Philippines when comparing their indigenous personality scales and the five-factor model, where they found

that some constructs were not accommodated by the model (Katigbak et al., 2002). As such, research towards the combination of etic test approaches (the cross-cultural application of personality tests in other contexts) and emic test approaches (the use of indigenously developed personality tests for that context) may produce a better understanding of both the universality and culturally specific personality traits (Cheung et al., 2011; Hill et al., 2021; Nel et al., 2012).

The development of the South African Personality Inventory (SAPI) not only adhered to the legislative guidelines but also considered the multicultural and multilinguistic considerations of the South African context (Hill et al., 2013, 2021; Nel et al., 2012). Much like the development of the Chinese Personality Assessment Inventory (CPAI), the research process for the SAPI relied on qualitative measures of obtaining lexical descriptions of personality and behaviour (Cheung et al., 2011; Hill et al., 2021). This involved the use of cluster analysis to reduce personality descriptors into 9 clusters used to understand personality within the different South African cultures (Cheung et al., 2011; Hill et al., 2013, 2021; Nel et al., 2012). Cheung et al. (2011) suggest that the projects mentioned above demonstrate cultural awareness and scientific rigour that is required for further development of these studies. This, therefore, leads the way for this study to employ computer analytical measures like text mining to better understand the formulation of personality clusters from linguistic data.

Factor analysis has led the way as a text clustering approach, which has been the foundation of constructing personality clusters (Péladeau & Davoodi, 2018). Some researchers have argued that it forms part of text mining. Text mining allows for the overall analysis of large chunks of data. As an algorithmically based computerised method, text mining introduces the aspect of mapping natural language usage at a larger scale (Antons et al., 2020). This method taps into stages such as text preprocessing, tokenisation, compound word identification, parts of speech relationship and semantic analysis. The use of such a method to analyse the use of personality traits within the South African context initiates the investigation into the validity of text mining approaches when understanding personality. Further, the data analysis method can explore the current personality inventories and their formulation and ascertain what the text mining approaches can understand from existing theories.

1.2 Rationale

Personality and personality traits often represent attitudes and people's behaviours with the influence of their cultural context (Matthews et al., 2009). The study of personality structures has yet to allow for an interpretation of the generalisability of traits due to the use of etic

approaches and foregoing the different experiences that different individuals have within different cultural settings (Morton et al., 2019). This means that the personality constructs are not adequately represented in different cultures, hence the development of the South African Personality Inventory (SAPI) (Cheung et al., 2011; Hill et al., 2013, 2021; Morton et al., 2019; Nel et al., 2012; Valchev et al., 2011). Using qualitative analyses and lexical strategies, nine personality clusters were yielded (Cheung et al., 2011; Hill et al., 2013, 2021; Nel et al., 2012). The personality clusters include; emotional stability, conscientiousness, facilitating, extraversion, integrity, intellect, openness, soft-heartedness and relationship harmony (Hill et al., 2013, 2021; Nel et al., 2012; Valchev et al., 2011). These represent the cultural relevance and the universal aspect of trait generalisability.

Researchers have shown the value of language as a medium of understanding personality. Schwartz et al. (2013) outline that social sciences have been leveraging the methods and resources that the digital and data science age has presented. This is done through computerised text analyses which are noted in the mining of written language from social media platforms (Nguyen et al., 2020; O'Connor et al., 2011; Schwartz et al., 2013). Such as cultural trend observation through quantitative word tracking of digitised books (Michel et al., 2011; Schwartz et al., 2013). The advancements towards Generative AI mean that psychology can tap into the use of Large Language Models (LLMs), which can help not only standardise psychological measures but also assist in the development of psychological constructs using existing linguistics output (Demszky et al., 2023). As such, the use of computerised and algorithmically based text analysis has been seen to draw on techniques such as machine learning and text mining to better comprehend human behaviour and society (O'Connor et al., 2011).

Thus, this study employs both unsupervised and supervised text mining techniques to understand the formulation of the personality structures using lexical personality data collected in South African literary text. These included utilising topic modelling, a probabilistic method of natural language clustering of topics within different documents (Silge & Robinson, 2017). Using such a methodology offers a comparative qualitative quantitatively-based perspective (O'Connor et al., 2011) for personality trait theory as it takes a step back in looking at the personality constructs derived from the traditional qualitative analyses. These come from the same linguistic data, which resulted in the personality model underlying the development of the SAPI. Further, using the Quanteda package for its function to perform analyses, such as co-occurrence, shows the importance of using emic-etic

approaches in understanding the structure of personality across cultures (Celardo & Everett, 2020). Using an integrative methodology of quantifying qualitative data, the study leverages data science methods to form a multidisciplinary interpretation.

1.3 Research Aims

The research aimed to explore personality structures found in South African literary texts. To understand how personality is portrayed in the South African context. This involved looking at a select list of South African literary texts using text-mining methods such as topic modelling. The development of the current personality classifier in South Africa SAPI utilised traditional qualitative methods to yield the existing clusters. This research aimed at not only finding the existing classification clusters using computational analyses but also looking at whether there is a difference between the existing personality clusters and those found using text mining techniques. This should further ascertain the validity of using computerised methods in understanding personality construction in South Africa.

1.4 Study Outline

This research paper presents five chapters that entail information regarding the exploration of text mining techniques to classify and understand personality traits and expressions. Chapter 1 outlines the background knowledge around personality research and the research methods currently used, providing a rationale for the inquiry into newer methods of investigation and the research aims. Chapter 2 lays out the literature on personality understanding, models, assessment, and relationship with language. This introduces the comprehension of trait theory, the methods of trait classification, the application of these methods, text mining and its comparison with factor analysis. Chapter 3 presents the methods used in the investigation, the type of design, processes, and analytical methods. Chapter 4 reports on the results, starting with a descriptive understanding of the text, the understanding of trait distribution and the use of unsupervised and supervised topic modelling. Chapter 5 discusses the results, locating them in existing literature and outlining the implications, limitations, and future recommendations for improving the study. Lastly, the conclusion presents the overall implications of text mining in exploring personality trait understanding and the need for further research on applying such a method.

Chapter 2 – Literature Review

2.1 Introduction

This chapter outlines the comprehension of personality from different theoretical perspectives. It further brings an understanding of the development of trait descriptions of personality and the linguistic contribution to trait personality theory. This led to a discussion on the development of personality models and their use within one cultural setting and cross-culturally. The literature review concludes by exploring the personality structure currently used in the SAPI scale and bringing understanding into text mining.

2.2 Personality

The search for understanding the unique differences among human beings has been an ongoing journey. Personality can be understood as a building block of interacting and comprehending one's environment (Kavirayani, 2018). It can be further seen as accurately representing oneself, attitudes, and behaviour. With no singular definition of personality, researchers have managed to explain its comprehension with trait descriptors (Matthews et al., 2009). The origins of personality understanding are derived from otherworldly beliefs, such as the Indian texts Vedas and the Bhagavad Gita, which attribute human nature and behaviour to being either demonic or divine (Kavirayani, 2018). Further, the belief in the constellation through astrology on how the positioning of the stars and time of birth attribute certain personality traits (Andersson et al., 2022; Chico & Lorenzo-Seva, 2006; Kavirayani, 2018). For example, people born under the star sign Aquarius born between (20 January – 18 February) are said to be independent, optimistic, and creative.

In the 5th century BC, a psycho-biological understanding of personality was presented by the Greek philosopher and physician Hippocrates (Eysenck, 1953; Kavirayani, 2018; Matthews et al., 2009; Schultz & Schultz, 2004). Using Galen's theory of humours, Hippocrates argued that bodily fluids such as bile and phlegm affect an individual's personality. The humours formed the temperament classification, which included sanguine, phlegmatic, choleric, and melancholic (Kavirayani, 2018; Matthews et al., 2009; Schultz & Schultz, 2004). Each of the temperamental classifications represents a psychological state, such as sanguine being connected to confidence and optimism, while choleric is connected to being hot-tempered (Kavirayani, 2018; Matthews et al., 2009). It was argued that an imbalance with these fluids leads to physical illness, thus resulting in mental disturbance (Matthews et al., 2009). Other theorists, such as William Sheldon, hypothesised that body shape can be associated with temperament (Kavirayani, 2018). Through this theory, the three trait classification was

formed: endomorphic (soft-appearing bodies, more sociable and easy-going), mesomorphs (muscular bodies, more assertive and outgoing) and ectomorphs (thinly appearing bodies, more introverted and artistic) (Kavirayani, 2018).

The introduction of the duality of body and mind by philosopher Rene Descartes (Kavirayani, 2018; Schultz & Schultz, 2004) meant the focus on understanding personality shifted towards the functioning of the mind. Thus, psychoanalysis theory, from its formulation by Sigmund Freud, presented that the unconscious factors influence the way an individual finds a balance (ego) in mitigating their personal drives (Id) and their social expectation (superego) (Kavirayani, 2018; Schultz & Schultz, 2016). Drawing from that, analytical psychology by Carl Jung first presented the classification of Extroversion and Introversion (Kavirayani, 2018). He argued that though people can be classified according to these broad categories, basic psychological functioning also needs to be considered (Jung & Hull, 1976; Kavirayani, 2018), which helps determine individual differences. These functions include thinking, feeling, sensing and intuition (Jung & Hull, 1976; Kavirayani, 2018).

2.3 Personality traits and language

Early modern research has hypothesised that the differences among people can be represented using natural language terms (De Raad & Mlacic, 2015; John & Robins, 2021; Matthews et al., 2009). The lexical approach outlines that personal traits of importance to individuals will be represented in their language use (De Raad & Mlacic, 2015). This means that if individuals consider themselves to be good people, their manner of speech will indicate what they consider about themselves. This is because conversations form part of one's behaviour and representation of others (De Raad & Mlacic, 2015). Uher (2013) outlines that the perception of personality traits is based on meaningful recurring patterns that can predict future events. These emphasise the consistency and stability of these individual differences (Diener et al., 2019). As such, these individual differences are encoded in language use and forming social relevancy in the portrayal of an individual (Goldberg, 1993; Uher, 2013).

2.3.1 Researcher-based trait collection

Personality investigation has involved looking into written assignments, interviews, and recordings to discern the use of language and its association with personality (Yarkoni, 2010). Word usage, therefore, relates to descriptors that form part of personality traits (De Raad & Mlacic, 2015; Goldberg, 1993; Yarkoni, 2010). Further understanding of the formulation of the linguistic attributes of personality follows a psycholexical approach. This involves the use of a dictionary to obtain words that are representational lexicons for the

abstract presentation of personality structures (De Raad & Mlacic, 2015; Matthews et al., 2009; Saucier, 2008). These words not only represent the personality characteristics but also the individuals' drive, dispositions, and orientation through their expression (Gunter, 2019).

Notably, the exploration of new languages means that certain traits will receive more attention while others receive less. In a study looking into personality-descriptive adjectives in Lithuanian, the Lithuanian-English dictionary was used to obtain 50,000 entries (Livaniene & De Raad, 2017). These words are assessed based on relevancy towards personality description and having exclusion criteria such as the reference to a part of an individual or metaphorical reference. A frequency of use Likert-type scale with 1 (this word is never used to describe a person) and 5 (highly used for personality description) (Livaniene & De Raad, 2017). This resulted in a final list of 435 personality-relevant adjectives.

In comparison to a study done on the trait lexicon for Hindi, (Singh et al., 2013), their study looked at descriptive adjectives for personality in Hindi through thesaurus as well as five Hindi novels for adjectives that could not be found in the thesaurus. A collection of 160 850 lexical entries was found in the thesaurus and 700 adjectives from the novels (Singh et al., 2013). An exclusion criterion was implemented, as well as a scale on the appropriate use of the descriptor for an individual's personality referred to as the appropriateness 3-point scale with 1 (least appropriate) and 3 (most appropriate) (Singh et al., 2013). This resulted in a final list of 295 personality-describing adjectives. Both studies emphasise the cross-cultural influence on personality. Livaniene and De Raad (2017) further highlight the use of fewer judges than the standard psycho-lexical studies, indicating that they could have had a longer list of relevantly describing words.

2.3.2 Computer-based trait collection

Yarkoni (2010) outlines that there are limitations to the current personality and language studies. The first constraint would be the collection of writing samples under study settings, limiting the natural occurrence of speech. Secondly, the sample sizes are often small and lastly, the broad view of the association between language and personality (Yarkoni, 2010). The use of computerised methods is a way to curb some of the limitations shown within the personality and language research. Currently, the frequency of interactions people have on social media has increased due to the everyday use of technology and the more accessible means of communication and social engagement.

As such, natural language use can be observed on these platforms (Christian et al., 2021a; Gunter, 2019; Schwartz et al., 2013). The extensive textual data provided by platforms like Facebook has offered a means of extracting linguistic qualities to gather insights into people's characters (Gunter, 2019). Studies like Golbeck et al. (2011) and Jeremy et al. (2019) investigated personality prediction on social media using language usage and profile representation. It was found that the user's behaviour was determined by the linguistic features that were found on their posts. These studies further yielded the advances made with the computerised methods used to understand personality traits using texts and their foundation in personality models (which would be later explored).

Computerised methods of human behaviour analysis have advanced the use of machine learning methods like Natural Language Processing (NLP) to detect personality traits in the text (Thisarani, 2021). This is through the mining of text recognised as part of the sentiment expression of human traits such as positivity and negativity and a range of other expressions from anger to trust and delight (Thisarani, 2021). This is to help understand individual thoughts, actions, and experiences. Numerous algorithms currently exist as extraction methods to understand personality embedded in social media posts, some of which are combined for better predictive outcomes (Christian et al., 2021b). While they use personality models (organised personality questionnaires) as a foundation for understanding individual personality, lexical databases have been formed using methods such as Linguistic Inquiry and Word Count (LIWC), Term Frequency-Inverse Document Frequency (TF-IDF), which is a closed-vocabulary extraction method to understand the relation between keywords on social media statuses and personality (Christian et al., 2021b).

The LIWC is a widely acknowledged application method of text extraction used to study personality and social understanding through linguistical features (Schwartz et al., 2013; Tadesse et al., 2018). Schwartz et al. (2013) present that the LIWC includes 64 language variables from part-of-speech to topic-based classification. The current version has 90 language variables with further expansion on psychological constructs (Pennebaker et al., 2015). A study by Pennebaker and King (2000) extracted text from different textual domains such as diaries, assignments, and social psychology manuscripts. It was found that neurotic individuals used more negative words, while agreeable individuals used more articles (Pennebaker & King, 2000; Schwartz et al., 2013). While adjectives have been used to represent personality traits, De Raad and Mlacic (2015) argue that this not only causes bias for psychological measures, but it does not represent the aim of the psycholexical approach as

it purports to offer a complete representation of a individuals' personality structure. The LIWC method manages to use a variation of part of speech in an attempt to offer a comprehensible understanding of an individual. Its application across different literary texts brings emphasis to its capabilities as a plausible method for personality trait research.

2.4 Personality models

The understanding of the formulation of traditional psychological assessment can be derived from two core aspects: (1) its reliance on the use of language and (2) a form of intervention in an individual's life (Gunter, 2019). This involves the textual observation of language use in either interviews done with respondents or the collection of questionnaires. These methods follow the lexical approach of researchers, pre-structuring the manner of response and providing guidelines on the words that can be used to answer the given questions (Gunter, 2019). However, a process of variable selection needs to be done. This involves understanding and finding linguistic representations of personality traits (Saucier, 2008). Arguably, linguistic correspondence must be significant in daily interactions, worldly represented and form part of "consequential criteria" (Matthews et al., 2009, p. 14). The consequential criteria indicate experiential significance, such as determining clinical disorders or measuring job performance and suitability (Matthews et al., 2009). These linguistic attributes will, therefore, form part of the personality structural model (Gunter, 2019; Matthews et al., 2009; Saucier, 2008).

The scientific method of factor analysis has traditionally been used to reduce the lexical representation into a single abstract attribute such as Extraversion or Introversion (De Raad & Mlacic, 2015; Gunter, 2019; John & Robins, 2021; Saucier, 2008; Schultz & Schultz, 2016). The origin of such an understanding of personality was cemented by Francis Galton. He proposed the lexical hypothesis, which states that human expressions and social engagements will eventually be represented by a single term across different languages (Goldberg, 1993; Roivainen, 2013). This method includes the use of a dictionary to extract personality descriptors contained in a lexicon and put them into categorisation (De Raad & Mlacic, 2015; Goldberg, 1993; Roivainen, 2013).

Gordon Allport explored the commonalities within human behaviour in developing the trait theory (Doremus, 2020). Trait theory aimed at formulating common traits to understand personality. In forming the groundwork for trait theory, Allport and his colleague Henry Odbert began by researching personality-describing words. Their work involved an inductive search for these descriptive words in the English dictionary (Allport & Odbert, 1936;

Doremus, 2020). Allport and Odbert (1936) initiated the use of scientific methods such as factor analysis to look through personality lexical terms in the dictionary and curate them into categorised personality traits.

In their study, the researchers came up with 17 953 descriptors, which were sectioned into four categories. The first category contains personality trait names; the second category contains state-of-being descriptors for both mind and mood; the third category contains character evaluations, and the last one contains physical descriptors, which are said to contribute to personality description (Allport & Odbert, 1936; Goldberg, 1993; Mollaret, 2009; Roivainen, 2013). It has been observed that the first and second categories are often used to represent an individual's personality. Mollaret (2009) argues that the personality trait categorisation performed by Allport and Odbert formed a basis for scientific personality classification.

Hans Eysenck came up with the three-factor model. Through factor analysis, he represented personality structure into three single-termed dimensions, namely, extraversion (extroversion vs introversion), neuroticism (negative emotionality), and psychoticism (disinhibition) (De Raad & Mlacic, 2015; Matthews et al., 2009; Saucier, 2008; Tohver, 2020). Tohver (2020) presents that the development of the three-factor model is supported by biological bases such as the neurological functioning of individuals. It can be further noted that the model can predict behavioural and lifespan outcomes. Eysenck and Eysenck (1991) present that an individual who scores high on the extraversion factor indicates that they are more positive and sociable. A high score on the neuroticism factor is indicative of emotional distress and anxiety feelings, while a high score on psychoticism shows an inclination to solitude, aggression, and lack of empathy (Eysenck, 1991; Matthews et al., 2009; Tohver, 2020). The neuroticism and the psychoticism factors may not be representative of everyday personality attributes, but they are acknowledged as being part of a standard personality classification (Eysenck, 1991).

The method of investigation of individual differences employed by Eysenck is theoretically based. This has raised criticism of the three factors being unable to adequately predict aspects of behaviour that can be observable (Boyle et al., 2016). Empirical evidence supports that a more significant prediction of individual differences is more likely to occur with a larger number of primary factors than with a smaller number of broad secondary factors (Boyle et al., 2016; Cattell & Krug, 1986). This led to the definition of personality that Raymond

Cattell cemented in studying human behaviour. He defined it as predicting what an individual will do in a given context (Cattell, 1943; Malkappagol, 2018).

Cattell, therefore, continued with the scientific centralisation of personality research. In understanding the empirical personality taxonomy, Cattell (1943) suggests that the measurement of personality constructs should involve 3 data categories: first, individual data collected in a real-life interaction (L data), second, data that is collected using assessments (Q data), and third, data from tests designed to measure actual behaviour (T data) (Cattell, 1965; Gillis & Boyle, 2018). This led him to follow the analysis and categorisation method for traits using factor analysis. The work allowed for the observation of the interrelation of traits to form higher-order factors (Cattell, 1965; Gillis & Boyle, 2018; Malkappagol, 2018). Resulting in the categorisation of 16 high-definition primary constructs (warmth, reasoning, emotional stability, dominance, liveliness, conscientiousness, social boldness, sensitivity, vigilance, abstractedness, privateness, apprehension, openness to change, self-reliance, perfectionism, tension) (Cattell & Cattell, 1995; Cattell & Krug, 1986; Gillis & Boyle, 2018) and five broad secondary constructs (anxiety, extraversion, independence, tough poise, control) (Cattell & Cattell, 1995; Gillis & Boyle, 2018).

The exploration into understanding personality structures led to the development of the widely known five-factor model. The Five-Factor Model (FFM) was proposed by McCrae and Costa in 1987 following the understanding behind trait theory that individuals can be characterised by their behavioural patterns, thoughts and feelings (Costa & McCrae, 1999). With the use of both the lexical strategies and the intervening text collected, five factors' representations of personality were yielded (Costa & McCrae, 1999; McCrae & John, 1992; Pincus, 2010). The model is laid out with five basic descriptions, which are classified as tendencies. These are Neuroticism, Extraversion, Openness to Experience, Agreeableness and Conscientiousness. The significant factors are supported by factor definers, which are taken from the natural language expression of personality (McCrae & John, 1992). It has been recognised that this personality model does not align with the lack of a specific factor from an individual's behaviour but rather the variation in the lower prevalence of that factor as opposed to another (Costa & McCrae, 2012; McCrae & John, 1992). As such, an individual cannot be classified as lacking in that factor but rather having a lower or higher inclination to it.

Each basic tendency is described by a characteristic adaptation, which explains the expected personality characterisation if an individual has high scores on that factor (Costa & McCrae, 2012; Pincus, 2010). This model's lexical representation of personality emphasises the individual and their state of being with little consideration of outside elements (Costa & McCrae, 2012). This is further brought into consideration by personality research in South Africa and China. Laher (2015) outlines that the FFM fails to draw in the interdependency personality has on culture, especially when looking at contexts like China. This is highlighted as one of the significant limitations of FFM, especially regarding the generalisability of the model and understanding and assessing personality differences cross-culturally (Cheung et al., 2011; Laher, 2015; Pincus, 2010).

The FFM is the most common model of understanding personality structure, which further brings an understanding of behavioural patterns and predicts future choices (John & Robins, 2021; Matthews et al., 2009). A study by Gurven et al. (2014) looked at the universality of the FFM among Bolivian forage farm workers, and it was found that there were socioecological characteristics that are often found in small communities. Thus, they have personality attributes related to their communal living (Gurven et al., 2013). Other studies looked at a comparison between Filipino personality scales and the FFM, and it was found that some personality constructs were accounted for. However, the culturally exploring constructs were not adequately represented in the model (Katigbak et al., 2002). The missing link between culturally based tests and cross-cultural ones is the relationality with the contexts. As such, further exploration needs to investigate more than the universality of traits; it needs to investigate the contextual representation as well. The current cross-cultural studies emphasise leading indigenous studies on personality and comparing them to the existing Western personality trait inventories (Thalmayer et al., 2021).

2.5 Personality Assessment

Researchers have looked into ways of measuring human behaviour. As such, personality assessment has offered a gateway to better quantifying and interpreting human behaviour. Ciccarelli (2006) mentions that personality assessments predict an individual's characteristics that remain constant and stable. This informs individual traits, what the person can do and what they are like, and further speaks to psychological treatment (Archer & Smith, 2011). Handler and Meyer (as cited (in Archer & Smith, 2011)) outline that the underlying premise of psychological assessment is not about getting a score like testing but combining contextual information and behavioural observation to understand the individual being evaluated. The

purpose of these assessments varies depending on the objective outcome, such as diagnostic purposes (clinical and organisational settings), individual understanding of personality using inventories and informing psychological treatment (Archer & Smith, 2011). The approach involves the use of either of the two types of personality assessment methods: projective and objective (Ciccarelli, 2006; Weiner & Greene, 2017).

Each assessment method leans in on different approaches to understanding human behaviour. The projective assessment method focuses on the use of objects that can be ambiguous or relevant to the individual in eliciting a response that is subjectively influenced (Ciccarelli, 2006; Reynolds et al., 2021). An example that can be drawn from this method is the Rorschach test, whereby an inkblot is used to discern an individual's personality and emotional functioning based on their observed response (Reynolds et al., 2021). Projective personality assessments rely on an individual's response, which is their subjective experience, which, unfortunately, can be filtered following desirability. Henceforth, the scientific basis of personality assessment brought about objective personality assessment, defined as tools designed with response items (True/False or Likert-type scales) that are scored objectively (Reynolds et al., 2021). These are, therefore, not influenced by individual judgement as they scored systematically. While the responses are scored systematically, it does not entirely remove the desirability responses but does mitigate the subjectivity behind them.

The construction of objective personality assessments is formed behind personality trait models such as the FFM. Tools like 16 Personality Factors (16PF), Myers-Briggs Type Indicator (MBTI) and Revised NEO Personality Inventory assess the presentation of individual personality traits in a scientifically constructed manner. However, the universal application of the assessments has been called into question, mainly because the traits used were constructed using WEIRD populations (Varadwaj, 2018). McCrae and Costa (1997) acknowledge that the five factors are not equally accessible in every context due to being influenced by genetically endogenous traits. Their study further notes that the prevalence of five factors in different cultures does not mean they are interpreted similarly (McCrae & Costa, 1997). Researchers have further noted that using assessments like the NEO inventory raises issues of reliability and replicability in other contexts, as such, calling for evidentiary indigenous collaboration (Thalmayer et al., 2022).

2.6 Application of personality theory and models: Recent research

The current theories of personality place emphasis on two aspects that are important in personality comprehension: genetic influence (McCrae & Costa, 1997) and language (Saucier, 2008). These have formed the congruity in understanding the natural language usage, contextual adaptation, and cultural influence towards forming personality and continual behavioural patterns. This can be seen in a recent study by Roivainen (2022), which attempted to understand personality terms within the Age of Acquisition (AoA) and their implications for personality.

The AoA is a psycholinguistic understanding of language acquisition in children and the influence of this process towards linguistic formation and adaptation (Roivainen, 2022). The study found that the initial words acquired by children (2nd grade or younger) are linked to personality words that form the core of the personality factors under the FFM as opposed to peripheral traits acquired later (Roivainen, 2022). Personality descriptors that had a high frequency of use were also found to have a high desirability rating (Roivainen, 2022; Wood, 2015). This indicates that the words that are socially favourable and gather social favour are learnt much earlier and tend to be used more than those that do not induce social desirability.

The understanding of personality traits and their expression in any context is a learned process. Personality theories have become fundamental in deciphering behavioural patterns and states of mind. Researchers introduce Personality States as an understanding of thought, feeling and affect in a given contextual space (Baumert et al., 2017). This extends to being the consistent expression of personality traits at a particular time. The understanding of such expressions helps articulate the involvement of personality traits in things like loneliness, leadership and social interactions (Buecker et al., 2020).

A meta-analysis study on loneliness and personality has indicated that individuals who are more agreeable, conscientious and open to experience are less likely to experience loneliness compared to individuals who are neurotic and introverted (Buecker et al., 2020). This is due to the positive correlation between extraversion and loneliness when other personality factors are controlled for leading to a heighten state of introversion, and individuals who are neurotic have a heightened sense towards social rejection (Denissen & Perke, 2008, as cited in Buecker et al., 2020). On the other hand, trait activation theory highlights the existence of personality states within organisational contexts. This is indicated by the expression of certain personality traits following the workplace setting, such as being more agreeable (Tett et al., 2021). Researchers have established the role that personality plays in choice and

decision-making. Another meta-analysis investigated the relationship between personality and performance, where it was found that a personality trait like conscientiousness is associated with an inclination to better performance in academics compared to traits like extraversion and neuroticism (Zell & Lesick, 2022). Further, looking into the influence the FFM has on online gaming, it was found that gamers tend to be more neurotic and having high conscientiousness protects individuals from being absorbed by the online world and not participating in the outside world (Akbari et al., 2021).

The research expansion shows that personality models like the FFM and the Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness and Openness to Experience (HEXACO) have become the source and structure of personality understanding. They are now models that have been read and interpreted across various cultures to further the knowledge of trait development. Lexical studies such as the understanding of descriptive characters in the Khoekhegowab, a language spoken in certain Southern African regions (Parts of Namibia, Botswana and South Africa), use the big five and the big 6 to interpret the construction and development of personality traits (Thalmayer, Job, et al., 2021; Thalmayer, Saucier, et al., 2021). More cultural studies have aimed to draw a degree of comprehension towards behavioural patterns and the influence that culture, upbringing, and language have in their expression.

2.7 The Chinese Personality Assessment Inventory (CPAI)

There is a difference in the manner of personality expression between either individualistic or collectivist communities. Individualism is considered a social pattern that prioritises an individual and their personal goals, and the individual is considered independent of the group (Triandis, 2018). While with collectivism, the individual exists within the group, their duties and goals are interconnected to that of the group. Triandis (2001) outlines that individuals who belong to more collectivist communities tend to align more with the group, with the focus of self-aligning going with that of the group and in consideration of the group and not only of the self. This was fundamental in the development of the Chinese Personality Assessment Inventory (CPAI) following the initial adaptation of Western-designed personality assessment tools like the MMPI (Cheung & Leung, 1998).

While the most common thing to do is to use translated Western personality tools in other contexts due to their pre-established reliability with supported conceptual and psychometric validity, the issue of construct interpretation arises (Cheung & Leung, 1998). The construction of the CPAI followed the misrepresentation of a cultural-comparable trait within

the Chinese community. As such, it introduced a cultural element in understanding personality formulation, development, and retention. Ko's Mental Health Questionnaire (KMHQ) was the beginning of the construction and understanding of Chinese personality traits in response to the use of etic (western-adopted) tests (Cheung et al., 2003; Cheung & Leung, 1998). Later followed by the Multi Trait Personality Inventory, which deduced that Chinese individuals from different contexts (China, Taiwan, USA and Hong Kong) had comparable traits that form part of the Chinese culture (Cheung et al., 1996a; Cheung & Leung, 1998; Cheung et al., 1992).

The Chinese Personality Assessment Inventory (CPAI) was, thus, developed with an established nature of the existence of Chinese culturally influenced personality traits. The personality constructs that came about had to be interwoven with the experiences of Chinese people and the linguistical interpretation of their daily lives (Cheung et al., 1996b). This led to the use of 5 methods to come up with personality constructs used in the Chinese context. First, about 15 popular contemporary Chinese novels written by Chinese authors from different contexts were selected for analysis. The books were filtered for adjectives used to describe the main characters (Cheung et al., 1996b; Cheung & Leung, 1998; Cheung et al., 1992). Second, a further investigation into the Chinese books of proverbs was done to better review some of the attitudes and behaviours used in describing individual characteristics. Third, researchers took an extra step by collecting 300 self-descriptive statements written by students in an attempt to understand and broaden the pool of trait descriptions (Cheung et al., 1996b). Fourth, various professionals also wrote a collection of statements to describe their colleagues with whom they regularly interact using ten adjectives. Fifth, an investigation into psychological literature examined the specific personality traits that manifested in Chinese communities (Cheung et al., 1996b). The traits include the face, ren-qing (relationship orientation) and somatisation.

The results of the various approaches to inventory construction include a total of 150 distinct personality characteristics. There was a 900-item scale with 26 typical personality constructs and 12 that were pathological and clinical (Cheung et al., 1996b). The CPAI was later revised in a trial in 1991, which ended up with a total of 22 typical personality constructs, with 12 being clinical-based and 1 being a validity scale. A standardisation trial in 1992 added two more validity scales with a 524-item scale (Cheung et al., 2013). The formulated constructs in understanding Chinese personality include normal and clinical constructs. The normal constructs include harmony, ren-qing, traditionalism-modernity, thrift-extravagance,

defensiveness/Ah-Q mentality, graciousness-meanness, voraciousness-slickness, face, family orientation, logical-affective orientation, flexibility, practical-mindedness, emotionality, responsibility, inferiority- self-assurance, optimism-pessimism, meticulousness, social sensitivity, internal-external locus of control, introversion-extraversion, leadership, adventurousness, relationship orientation, enterprise, interpersonal tolerance, aesthetics, divergent thinking, diversity, and novelty. Clinical constructs: depression, inferiority, physical symptoms, anxiety, somatisation, need for attention, hypomania, antisocial behaviour, paranoia, pathological dependence, distortion of reality and sexual problems (Cheung et al., 2003; Cheung & Leung, 1998).

In a comparative study against the five-factor model, it was found that there were no openness attributes in the CPAI constructs (Cheung et al., 2008). This resulted in a study that led to the extension into the CPAI-2, which included aspects related to the attribute of openness to experience related to the Chinese culture. The final inventory included openness-related attributes such as diversity, aesthetics, novelty, and divergent thinking (Cheung et al., 2013; Cheung et al., 2008). Four personality factors were extracted for the CPAI-2: accommodation, dependability, interpersonal relatedness, and social potency. The formulation of the CPAI to CPAI-2 encapsulated the cultural relatability aspect of personality and a universal understanding of personality through appropriate linguistical comprehension of daily life experiences.

2.8 Development of the South African Personality Inventory (SAPI)

Contextual experiences play a huge role in developing and standardising psychological assessments. In a country like South Africa, where history has influenced the creation of bias and unfair measures (Cooper, 2014; Laher & Cockcroft, 2014), the development of the SAPI had to consider different factors. This includes being guided by the Employment Equity Act 55 of 1998 legislation in ensuring that the measure is fair, valid, and reliable, acknowledging diverse backgrounds and rectifying historical ill-doings (Government Gazette, 1998).

Previous attempts at conceptualising personality in the African context involved using Western-based assessments and models (Valchev, 2012). Any local attempt to develop personality measures in South Africa showed minimal relevance within the different cultural groups and did not take into consideration the existence of different languages (Abrahams & Mauer, 1999; Nel et al., 2012). The development of this inventory puts forward four elements to be addressed: (1) important personality concepts in an integrated multicultural context like South Africa, (2) how personality concepts differ within main cultural groups, (3) the

interchangeable role played by trait and context and (4) the implications of contextually-based personality models when generalising the findings (Nel et al., 2012; Valchev, 2012). Valchev (2012) outlines that the investigation into individuals extends to their understanding and occupation of their culture.

The research on the development of the SAPI began with understanding implicit personality conception in Nguni cultural-linguistic groups, which were Swati, Zulu, and Xhosa (Valchev, 2012; Valchev et al., 2011). The five-factor model was used to develop construct equivalence with the universality aspect of the FFM and further verify the path of the obtained outcomes (Valchev et al., 2011). The study followed the lexical strategy of pre-structuring the interviews with words that could guide the personality description of the individual (Valchev, 2012; Valchev et al., 2011). The responses were collected, transcribed, and translated. The results indicated that personality conceptions are similar among the Nguni cultural-linguistic groups. Concluding with 26 clusters, this further indicated that the clusters obtained from the study are not represented in Western models like the FFM. A trait like guidance was notably common among the cultural-linguistic groups (Valchev, 2012; Valchev et al., 2011).

The project proceeded to study the personality structures among the 11 official languages (Zulu, Swati, Xhosa, Ndebele, English, Sotho, Pedi, Tswana, Venda, Tsonga, and Afrikaans) of South Africa and their representative cultures (Nel et al., 2012). The analysis of the study involved three parts: labelling, categorising and semantic clustering (Nel et al., 2012; Valchev, 2012). Nine clusters were obtained; these include conscientiousness, extraversion, emotional stability, facilitating, integrity, intellect, openness, relationship harmony, and soft-heartedness (Nel et al., 2012; Valchev, 2012). These personality structures can be seen to share similarities with the FFM.

A notably relational trait of Ubuntu is acknowledged as an existing element that brings an understanding of personality from a cultural and environmental context (Nel et al., 2012). This draws on the use of cultural-based and cross-cultural assessments of personality. A quantitative approach was also implemented by administering questionnaires. Using the hierarchical cluster analysis, 37 subclusters were reduced to get 9 cluster classifications (Nel et al., 2012; Valchev, 2012). The results from this project build a solid basis for developing a cultural-inclusive model and further create culturally appropriate assessments for the South African context. Thus, this study will use computerised data clustering methods through text

mining to not only try to replicate but further observe any new clustering that might occur because of the methodological approach.

2.9 Text mining

The inquiry and understanding of text have come to require more computerised methods because of their large quantity. The expansion towards social media studying, where sites like Facebook produce about 4 petabytes of data daily (Osman, 2022; Talib et al., 2016), suggests a more computerised method with more robust analyses is needed. Text mining is defined as the extraction of meaning and insight from any given text (Talib et al., 2016). The origins of text mining came from the field of information retrieval, data mining and natural language processing, with the focus being on text retrieval using computerised methods (Talib et al., 2016; Zanini & Dhawan, 2015). Text mining utilises processes such as information retrieval, natural language processing, information extraction and data mining.

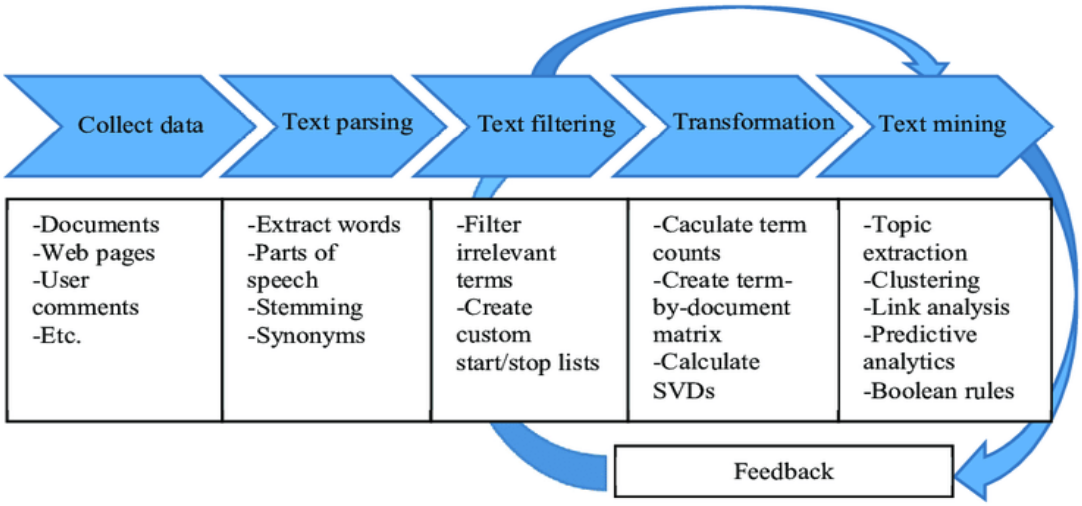


Figure 1. Text Mining Process. Source: (Kahya-Özyirmidokuz, 2014, p. 322)

2.9.1 Information retrieval

The information retrieval process is the initial technique involved in text mining. It includes the collection of textual data that will be used for analysis. This involves filtering information that will be relevant from those that will not be relevant (Babu et al., 2018). The information retrieval technique is concerned with organising large textual information documents. The documents can vary from legal documents, journal articles, and books to web pages, a compilation of Twitter (now X) posts and product reviews.

2.9.2 Text Parsing and Filtering

The second technique in text mining is text parsing and filtering, which involves sorting the textual data. This technique revolves around cleaning and parsing the textual data into formats that can be better interpreted and analysed (Wang et al., 2022). The analytical processes require structured data that removes any form of error, fixes typing mistakes and removes words that do not align with the language of analysis (Gaikwad et al., 2014). This process is referred to as normalisation, which refers to the transformation of words into uniform formats that are easier to analyse (Welbers et al., 2017). This technique includes removing stop words, which removes parts of speech that do not give insight when analysing, such as articles, auxiliary verbs and conjunctions (Gaikwad et al., 2014). The process of stemming entails the reduction of a word to its stem by removing any prefixes, suffixes, and tense changers. Lastly, part of speech tagging is the identification of words and the part of speech assigned to them.

2.9.3 Text transformation

The technique of textual transformation entails raw text manipulation whereby the text in the documents is transformed into matrices. These matrices are created by turning words into tokens, also called tokenisation. These are units of text that form a bag of words, and these texts are split into tokens (Welbers et al., 2017). The step of tokenisation is essential for analysis because full words are found to be too broad to produce any meaningful analyses. In languages like English, it has been found that performing tokenisation is easier because words are separated by punctuation and white spaces, whereas in languages like Japanese and Chinese words, dictionaries are required to define the separation of words into tokens (Welbers et al., 2017). Packages like quanteda in R using the tokens function make the process of tokenisation much easier.

The Document Term Matrix (DTM) is formed by the tokens as part of the bag of words. The matrix consists of rows that represent the documents and columns that represent the words (Welbers et al., 2017). The efficiency of the matrix being in this format comes with easily converting the words to numbers. In R, the representation of this can be done as the document term matrix (DTM), which is performed under the quanteda package, or the Term Document Matrix (TDM), which is done under the tm package (Silge & Robinson, 2017; Welbers et al., 2017). The matrices then form the foundation of any analytical and visualisation approaches.

2.9.4 Text analysis

The last technique of text mining consists of textual analysis and data visualisation. Boumans and Trilling (2016) outline that this technique consists of (1) counting and dictionary methods, (2) supervised learning methods and (3) unsupervised learning methods. The order of the processes within the analysis technique is aligned in the manner of one being the most deductive method to the most inductive method (Boumans & Trilling, 2016; Welbers et al., 2017). The deductive aspect focuses on defining the data through theoretical backing, which means the researchers know what they are looking for in their data. The inductive method draws its focus from the data and lets the data inform any findings (Bonner et al., 2021).

The counting and dictionary method, being more deductive, involves using predefined Boolean queries, expressions, and keywords to observe the pattern of occurrence of such predefined terms. The dictionary often contains these terms, which are compared to the primary text (Welbers et al., 2017). Furthermore, supervised learning involves inputting data into a model and pattern recognition is based on the data sample, which is trained, allowing the model to know which path to follow based on the predefined sample. It involves techniques like regression and classification. Unsupervised learning, being more inductive, has no predefined parameters, allowing for the formation of new and undiscovered data patterns (Hiran et al., 2021; Q. Liu & Wu, 2012). It involves techniques such as clustering and dimensionality reduction. The processes yield outcomes that allow researchers to visualise the findings and their applicability to any context.

The text analysis methods are used to understand the expression of sentiment in natural language processing. Sentiment analysis involves the analysis of emotions, attitudes, and opinions (Cheng & Chen, 2019). Often, sentiment analysis is used to understand consumer behaviour through social media posts and product reviews (Cheng & Chen, 2019). Text analysis extends to feature extraction involving the extraction of traits such as words of species names from journal article titles and abstracts in a biology study (Kaur et al., 2019). A study by Bonch-Osmolovskaya and Skorinkin (2017) on using text mining in the automatic extraction of lexical patterns associated with different characters in a novel to better understand literary techniques authors use in their character verbal distinction. The study shows that the analysis of syntactic dependencies has the potential to highlight character development and literary techniques (Bonch-Osmolovskaya & Skorinkin, 2017).

2.10 Topic Modelling versus Factor Analysis

The methods of analysing text aim to discern the underlying relationships and meaning in any given document. Hence, techniques such as factor analysis are used to understand meaning. Factor analysis is defined as a statistical procedure that simplifies variables/items to find a relationship between variables and underlying factors in each item (Wetzel, 2020). This technique has been fundamental in the development of personality clusters and the understanding of trait classification (Beck & Jackson, 2021; Gillis & Boyle, 2018). It has been used to form the currently used personality clusters by assessing factor loading, which is the correlation between the item and the factor. The approach contains aspects such as Exploratory Factor Analysis and Confirmatory Factor Analysis. Péladeau and Davoodi (2018) outline that factor analysis was the initial effort of text mining as it was used to automatically generate topics stored in the content analysis dictionaries. Its underlying feature is an unsupervised method of discovering latent variables, similar to Latent Dirichlet Allocation.

In their study comparing Latent Dirichlet Allocation, which is a technique under Topic Modelling and Factor Analysis in topic extraction, they define Topic Modelling as a probabilistic technique that looks at recurrent patterns of words in documents (Péladeau & Davoodi, 2018). Topic modelling uses Natural language Processing to understand the sentiment, retrieve information, and summarise the text. They outline that there is little literature capturing the improvement of text mining from FA to topic modelling (LDA) (Péladeau & Davoodi, 2018). This led to the objective of this study to explore the use of text mining techniques in understanding personality expression and trait classification. As the Dirichlet functionality looks, the pattern of words that repeat together occurs frequently, as well as similar words.

2.11 Research Questions

- How are personality descriptors classified using text mining?
- Are there different personality clusters produced using topic modelling in South African literary texts?
- What are the common themes found in the South African literary text in comparison to existing personality classifications?

2.12 Conclusion

The chapter managed to outline the historical and theoretical understanding of personality. Further, introduces a discussion on integrating language and understanding personality and personality traits. The development of personality models and their applications in different cultures. The research behind the development of the SAPI and the CPAI. Concluding with an understanding of what text mining is and its approaches.

Chapter 3 – Methods

3.1 Introduction

The methods chapter outlines the approach that the study used to answer the research questions and fulfil the aims. This includes the proposed study design, the data to be used, the procedure followed, and the ethical considerations. Lastly, the proposed data analysis process for the study will be presented.

3.2 Research Design

The study followed a document analysis. This process reviews and evaluates physical and electronic documents (Bowen, 2009). This form of analysis utilised the text-mining approach. This involves a qualitative algorithmically screening through the list of books in search of personality descriptors that could be presented as adjectives. This method looks through the concepts contained within each sentence by analysing the semantic structure of each sentence and then the whole document (Gaikwad et al., 2014). Text mining has two designs: supervised and unsupervised methods. The supervised design involves having predefined terms and checking for their representation in the list of text, while the unsupervised design relies on finding those terms in the text (Liu & Wu, 2012).

An unsupervised method was used to allow the algorithm to assign thematic topics following the most occurring trends. A semi-supervised method was applied, following the principles of a supervised method in guiding the process and the freedom of the unsupervised method in finding its associations. This process was guided using predefined personality descriptors. These are then compared to the literary text to investigate the commonalities with these themes. The choice for the text mining method was because of its ability to distinguish between significant and insignificant terms and relies on natural language processing (Gaikwad et al., 2014). The study followed a nomothetic approach to explore the existence of generalities within any given text. This will, therefore, be observed within the list of selected South African texts. This can assist in accounting for any personality trait commonalities and explain any differences (Ignatow & Mihalcea, 2017).

3.3 Sample

The data used consists of 60 South African novels that South African authors wrote based on South African stories (See Table 2). Novels as a textual genre were chosen because of the style of writing consisting of character descriptions. Further, previous personality construct research in China has used the method of extracting personality descriptors from novels (Cheung et al., 1996a) as characters are described following cultural relevance. The study

implemented a convenience sampling technique that included finding accessible text for analysis. These books were conveniently selected based on books that could be accessible in PDF format and from library resources. The books had to be in PDF format for appropriate analyses using the statistical tool. The list of books is made up of books written by well-known South African authors such as JM Coetzee, Zakes Mda and Nadine Gordimer. The books refer to various South African cultures, such as Afrikaans, Zulu, and Xhosa.

3.4 Procedure

The search for books was a bit of a challenge because few South African novels are accessible in PDF formats on library sites. The PDF format was a prerequisite in the book search because they are easier to analyse by the software tool. Scanning physical books will not only exceed the fair usage guidelines of the Copyright Infringement Act, but the tool will also be unable to analyse the text. As such, a librarian had to be contacted to request a list of available South African books in this format. Communication was back and forth through emails and various librarians. During communication, a list of South African books was requested based on South African authors known to have published books within diverse cultures. However, the request was not fruitful because the library could not find these books and could not order them because they were not part of the literature curriculum. The books that were procured were the ones that the library already had in that format. This resulted in a total of 60 novels that were analysed using text mining.

3.5 Ethical considerations

The study used publicly available books that are fictionally based, meaning no personal interaction or interaction with personal information. Thus, an ethical waiver was applied through an internal ethical clearance procedure within the E-Science program and the Department of Psychology under the University of the Witwatersrand Ethics Committee. The protocol number MAPSYC-22-04W was issued for ethical reference for the study.

The books are readily available pieces of text with cultural relevance but contain no personally identifying information, so there was no ethical infringement. Further, the books are not copied, indicating no copyright infringement concerning fair usage of the texts. There was an analysis of the books, using text mining, meaning no personal information is being used beyond the availability of the author's name. The process of appropriate usage and safekeeping of the data was followed and guided by the Protection of Personal Information Act (POPIA) (Government Gazette, 2013) as well as the Copyright Act (Government Gazette, 1978). This included keeping the data in a password-protected laptop and an

encrypted folder. Further, the text was fairly used, such as only looking into the personality descriptors in the books presented as adjectives. The data was subjected to text-mining analytic techniques.

3.6 Data Analysis

Text mining tools were used to analyse the books using the R programming language (R core Team, 2023), which is meant for statistical analyses, data manipulation, and data visualisation. Supervised analysis methods were used, namely the LDA, to list pre-existing personality descriptors found in the South African selected books. Furthermore, an unsupervised learning method of analysis was also used for the detection of unknown patterns in unlabelled text (Sathya & Abraham, 2013).

Text pre-processing steps were followed. These included corpus creation, tokenisation, and creation of a document feature matrix (DFM) (Antons et al., 2020; Ignatow & Mihalcea, 2017). The Quanteda package in R was paramount in developing and creating the corpus and document matrix and forming part of the analysis. It was used because of its versatility when it comes to text mining analytics (Benoit et al., 2018; Monroe, 2021). The method of analysis used included topic modelling, understood as a form of concept analysis that establishes the prevalence of a particular topic within and between documents (Silge & Robinson, 2017). Topic modelling contains facets such as the Latent Dirichlet Allocation (LDA), which allows for unsupervised and semi-supervised analyses to be performed. Other methods of analysis include Latent Semantic Analysis, Non-negative Matrix Factorization, and Parallel Latent Dirichlet Allocation.

3.6.1 Data Cleaning

Text mining analysis involves data pre-processing to ensure that the text is standardised and easier to analyse. The process followed the steps recommended by Kwartler (2017). To start, R packages that would be useful for the analyses were loaded and read in the PDF text file into the workspace. Four articles were loaded in to pilot the coding for textual analysis before analysing the complete textual list. The pre-processing routine involved changing the PDF text files to .txt files through the application of a function and locating the files in the device. This ought to make it simpler to read the contents of the books into the R workspace. However, there were difficulties with the file being converted. A function was then created that allowed the PDFs to be loaded and the content of each book to be analysed.

A corpus was therefore created consisting of all the contents of all the individual text files. Tokenisation is a process of changing individual words into tokens (Gaikwad et al., 2014;

Kwartler, 2017). The processing of tokenisation included screening the text and removing digital numbers, stop words which are non-informative (e.g. "I", "The"), common punctuation marks (e.g. ".", "?"), contracted words (e.g. "isn't", "don't"). One of the steps, stemming, was not implemented. Stemming involves the act of reducing a word to its stem, which removes any prefix or suffix added to the original word (Silge & Robinson, 2017). This process was not used as many words that align with personality descriptors and words presented in the dictionaries, including words with prefixes and suffixes. The text contained stop words that were not in English, such as Afrikaans and IsiZulu words. These were removed.

3.6.2 Coding Process

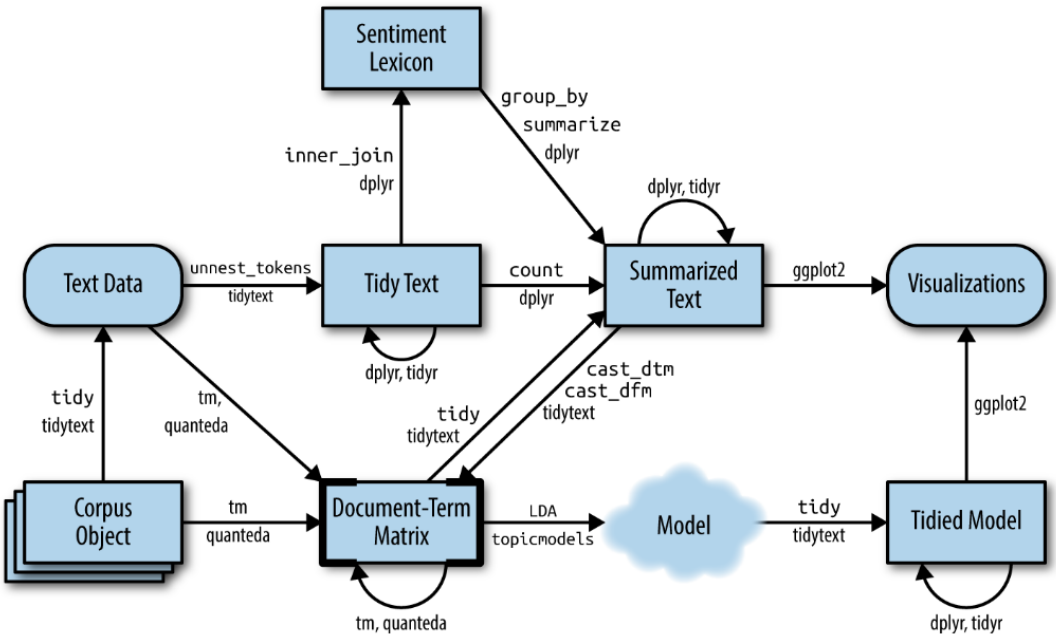


Figure 2. The data manipulation process. Source: (Silge & Robinson, 2017, p. 87)

Figure 2 illustrates the process and the various packages used for the processing, manipulation, and textual analyses. The first attempt included the documents being loaded individually. This posed a problem because the text needed to be analysed as one document. The text files were then entered into the R script as one, and a corpus was created. A corpus contains all the text files under a single chunk. These were cleaned, and a Document Term Matrix (DTM) was created, a matrix containing all the words from the files, excluding any stop words that have been cleaned. The DTM forms the basis of an analysis and is used to run any word frequencies, commonalities, and word clouds. The outcome summary includes texts that have been tokenised. The study focused on finding personality descriptors in South

African books, so a dictionary of personality descriptors was created containing personality descriptors from English, the Chinese Personality Inventory, and the South African Personality Inventory (SAPI). These were compared against the DTM, and an issue was found due to the misalignment of analytical packages. When constructing the DTM, a package called Tidytext formed the foundation, but the dictionary comparison uses a different foundational package called Quanteda (Benoit, 2018).

Quanteda was used for this kind of textual analysis because of its different analyses and the range of comparisons it can perform (Monroe, 2021). For the second attempt, the corpus was reconstructed using standard Quanteda packages. This involved the steps of text pre-processing before corpus creation, removal of digital numbers, web links, punctuation, stop words, making every letter small and removing words attached to punctuation. A Document Feature Matrix (DFM) is a matrix of all the words that have been tokenised; these words are not in a sentence base but are coded as singular words.

The approach of observing and understanding the personality descriptors began with topic modelling. Topic modelling is a process of classifying words into topics that indicate their likely occurrences (Kwartler, 2017). This process contains several techniques that identify words in many ways, such as the LDA, which is the technique that was used in this study involving the development of probabilistic representation of words (Blei, 2003). It was chosen because it can create topics that can classify personality descriptors and further compare any personality descriptors to those presented in dictionaries.

3.6.3 Analyses

The data analysis process involved running multiple forms of analyses to capture the contents of the text and gather meaningful insights relating to personality representation within South African literary text. The first analysis was to capture the content of the text by displaying the corpus and the contents of each book. The second analysis was to assess the readability of the text. The Flesch index uses the total number of sentences, the total number of words and the total number of syllables to produce the readability index (Talburt, 1985a). The document feature matrix was formed to categorise the words in a matrix format. A matrix of parts of speech was formed using parts-of-speech (POS) tagging, which classifies each word into a part of speech. This led to the classification of adjectives (Bonch-Osmolovskaya & Skorinkin, 2017; Gaikwad et al., 2014).

The frequency of the adjectives was observed, determining which are the most used adjectives in the books. An overall topic modelling was done to find any underlying topics in the text. This further led to the dictionary-based LDA, which categorised the first output as SAPI, English and Chinese, which were the personality descriptors formed. The second output was classified following Goldberg's 5F traits. An adjective categorisation was done to check the existence of personality-aligning traits. This involved looking at co-occurrences between pronouns and Goldberg's trait list. This led to constructing an adjective data frame, which was clustered using the LDA approach.

3.7 Conclusion

This chapter presented a methodological approach that includes the research design the study followed, the data that was used, the manner of collection, and the procedure that was followed in obtaining the data. Further, the ethical considerations that needed to be considered when doing the study and, lastly, the method of analysis that was used to yield results were discussed. The results obtained are presented in the next chapter.

Chapter 4 – Results

4.1 Introduction

The results section presents information on the classification of personality descriptors obtained from the literary text. The results are arranged under the questions that were asked regarding personality descriptors classified using text mining, personality structures classified using topic modelling, highlighting the different classifications compared to SAPI and discerning themes found in the literary text compared to existing personality classifications. It further shows parts-of-speech (POS) tagging, which involves the extraction of adjectives to discern their use as personality descriptors.

4.2 Descriptive textual analysis

This section conveys the foundation of the analysed textual data. The raw South African literary texts are broken down and presented as a corpus, which summarises the contents of the texts. Further, the understandability of the text is presented by outlining the level of readability. This supports the idea that any reader can understand any personality traits used. The frequency of the most occurring words was evaluated to gather further insight into the structure of the text and the relevancy it might have towards personality presentation.

4.2.1 Corpus

The analytical processes were done on a list of 60 South African literary texts, represented in Table 2 as a descriptive summary with complete information on the books. A summary of the Corpus presents the breakdown of the information contained in each of the books on the list of texts. Table 1 (see Appendix) presents information on the contents of the texts; this includes the characters of the text, the sentences, combined tokens, types, punctuation, numbers, symbols, website links (URLs), tags and emojis. The list of texts contains no emojis mainly because these are novels; thus, any emotional expression is worded instead of pictorial representation. The table below (Table 2) presents the summary of the analysed text, including the author's name, book title, publication year, tokens and the Flesch index.

Text	Author name	Book title	Publication date	Tokens	Flesch index
1	Deon Meyer	7 days	2012	135511	80
2	Alex La Guma	A walk in the night	1967	50421	82
3	Marlene Van Niekerk	Agaat	2004	282942	83
4	J.M Coetzee	Age of Iron	1990	70568	87
5	Jassy Mackenzie	Bad seeds	2017	112031	77
6	Miranda Sherry	Black dog summer	2014	100639	80
7	Malla Nunn	Blessed are the dead	2012	98807	75
8	Yemande Omotoso	Bom boys	2011	61132	82
9	Deon Meyer	Cobra	2014	131363	81
10	Kopano Matlwa	Coconut	2007	55412	79
11	Sheila Kohler	Cracks	1994	55105	77
12	Alan Paton	Cry the beloved country	1948	103142	81
13	JM Coetzee	Diary of a Bad Year	2007	68458	69
14	J.M Coetzee	Disgrace	1999	81837	83
15	J.M Coetzee	Dusklands	1982	57001	71
16	Nadine Gordimer	Get a life	2005	54524	61
17	Deon Meyer	Heart of the Hunter	2002	137105	77
18	Bianca Marais	Hum if you don't know the words	2017	143852	82
19	Bianca Marais	If you want to make God laugh	2019	130283	79
20	Tom Sharpe	Indecent Exposure	1973	97219	70
21	Nadine Gordimer	July's people	1981	55162	70
22	Deon Meyer	Koors	2016	208529	76
23	J.M Coetzee	Life and times of Michael K	1983	76890	82
24	Bessie Head	Maru	1971	42007	77
25	Tim Willocks	Memo from Turner	2018	116582	84
26	Wilbur Smith	Men of men	1981	256891	74
27	Mark Winkler	My name is Nathan Lucius	2015	72661	90
28	Deon Meyer	Orion	1998	141468	69
29	Jassy Mackenzie	Pale horses	2012	106580	77
30	Kopano Matlwa	Period Pain	2016	33435	80
31	Zoe Wicomb	Playing in the light	2006	93382	69
32	Jassy Mackenzie	Random Violence	2010	109965	83
33	Linzi Glass	Ruby Red	2008	64203	78
34	J.M Coetzee	Scenes from provincial life	1997	208726	76
35	JM Coetzee	Slow Man	2005	94404	81
36	Kopano Matlwa	Split Milk	2010	53513	79
37	Bryce Courtney	Tandia	1992	391891	74
38	Angela Makholwa	The blessed girl	2017	81969	81
39	JM Coetzee	The Childhood of Jesus	2013	103768	87
40	James A. Michener	The covenant	2015	561393	67
41	Ivan Vladislavić	The distance	2020	85535	77
42	Damon Galgut	The good doctor	2003	85383	84
43	Zakes Mda	The heart of redness	2000	116987	74

44	Deon Meyer	The last hun	2019	141421	78
45	JM Coetzee	The lives of animals	1999	48133	59
46	Dave Boling	The lost history of stars	2017	108536	86
47	J.M Coetzee	The school days of Jesus	2016	93110	82
48	James McClure	The song dog	1991	111100	76
49	SL Grey	The Ward	2012	115827	86
50	Zakes Mda	The whale caller	2005	89621	79
51	Deon Meyer	The woman in the blue cloak	2017	38557	77
52	K Sello Duiker	Thirteen Cents	2000	63341	89
53	JL Powers	This thing called the future	2011	75854	85
54	Zakes Mda	Ways of dying	1995	79607	77
55	Karin Brynard	Weeping Water	2009	163076	84
56	Eleanor Morse	White dog fell from the sky	2013	140605	84
57	Bryce Courtney	Whitethorn	2006	289910	75
58	Zoe Wicomb	You can get lost in Cape Town	1987	80299	70
59	Damian Barr	You will be safe	2019	89453	81
60	Sifiso Mzobe	Young blood	2010	70789	62

Table 2. Text Summary

4.2.1.2 Textual Readability

A readability test was done to discern the readability rate of the different texts. The readability index represents "the average number of syllables per word and the average number of words per sentence" (Talburt, 1985, p. 114). The Flesch index is used to quantify the level of readability and understanding of the text (Talburt, 1985). The table above (Table 2) presents the Flesch index of the literary text.

The Flesch index exists within a spectrum from 0-100. This is to better understand the level of readability of the text. Talburt (1985). A range between 0-30 indicates that the text can only be read by college graduates, between 50-60 indicates that high school graduates can read the text, and the 90-100 level indicates fourth-grade learner readability (Talburt, 1985). The Flesch index in Table 2 indicates that an average adult can read the text as the lowest index presented above in text 45 with a Flesch index of 59. The readability score is within the 50-60 range, indicating that the terms used to describe characters can be easily understood; as such, the personality descriptors are easily communicable.

4.2.1.3 Word Frequency

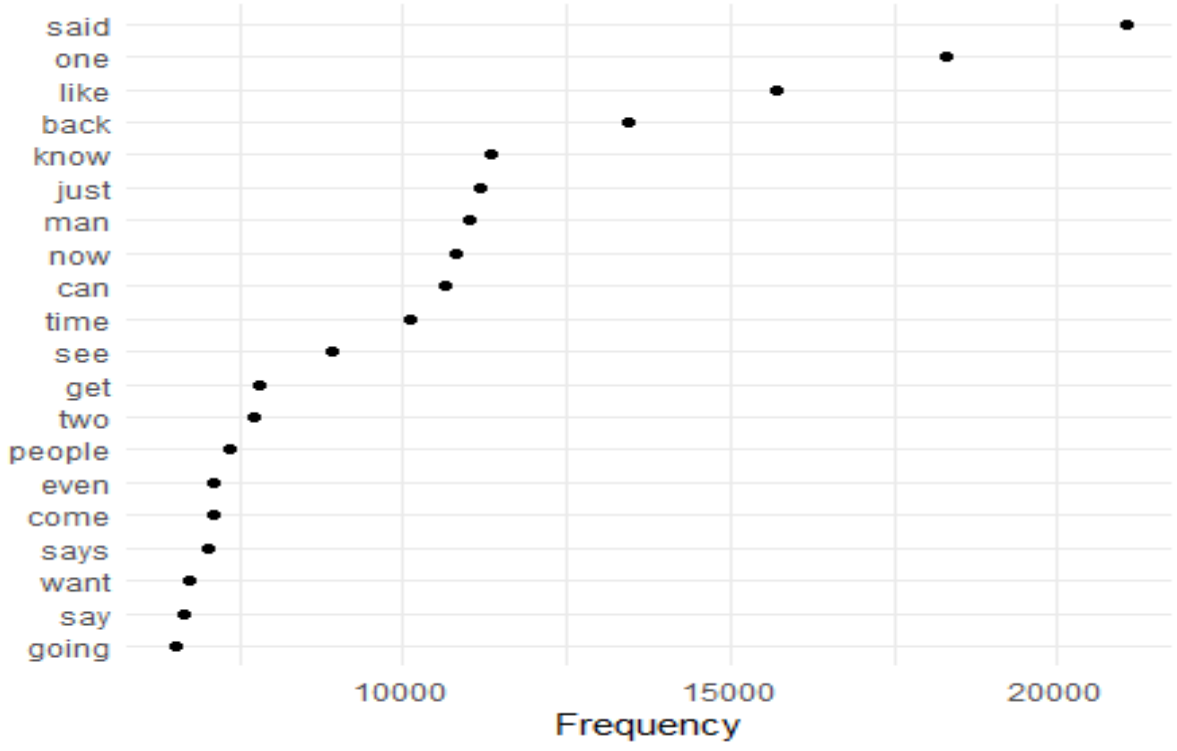


Figure 3. Frequency Distribution

Figure 3 presents the document word frequency of the top 20 most frequent words in the list of texts. The data indicates the most common feature and its frequency of occurrence across the different textual documents. The figure illustrates the frequency distribution of the most common features in the text. The most used word is *"said"* with a frequency of 21041. The reason behind this could be that these are literary texts; as such, the narration of character stories is the approach in this type of text. The frequency of words does not give a depth of insight relating to the personality description of characters because these would be words that are not frequently used in literary texts.

4.2.1.4 Parts of speech distribution - Adjectives

The textual properties of the books involve understanding the contents and distribution of the adjectives that would form part of the personality descriptors, as well as the distribution of parts of speech and any co-occurrences of adjectives and nouns. Using the *UDPipe* package (Wijffels et al, 2020), the distribution of adjectives in the text was done. The graph below shows the frequency of the most occurring adjectives in the text.

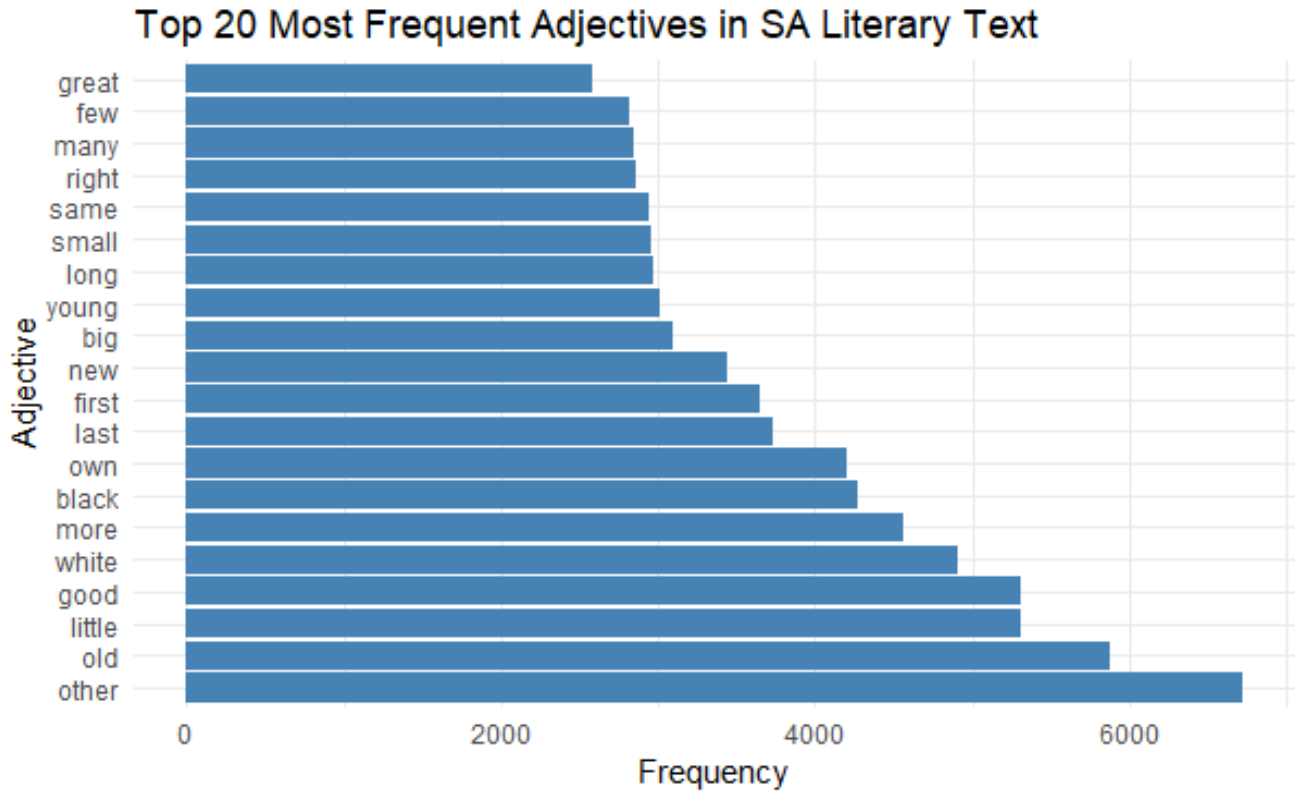


Figure 4: Top 20 most occurring adjectives in SA text

Figure 4 presents the commonly used adjectives to describe individuals, places, or objects. Above are the 20 top selected adjectives used as descriptors on the text that appear at least once in each given textual writing. A couple of the above adjectives correspond to personality descriptors such as *"good"*. These, however, do not give enough context and link to them being solely used as personality descriptors. The adjectives that are commonly used are often used in the context of referring to other things, such as words such as *"old"*, referring to a state of being and *"first"*, referring to positionality (Grzybowski et al., 2021). These descriptions do not refer to an individual and their behavioural manner. This indicates a further need for filtration of descriptors that speak to understanding an individual and their consistent behavioural traits.

4.3 Personality trait distribution

4.3.1 Trait variation in different gender identities

With the establishment of all the adjectives and the most frequently used adjectives, a link between adjectives that are personality descriptors and pronouns was observed. Using Goldberg's personality list (Goldberg, 1993), a list of pronouns was constructed to find the co-occurrence between adjectives and pronouns. This was done to discern the use of person-describing adjectives in the literary text. Goldberg's list was used because not only does it contain a smaller number of trait descriptors, but the traits are also attributed to their universal application through the FFM. The pronoun grouping involved words and pronouns that refer to each gender or sex. In this case, male pronouns were *("he", "his", "man", "Mr", "boy"*) and female pronouns were ("*she", "her", "woman", "miss", "Mrs", "lady", "girl"*).

This pattern indicates the use of gendered adjectives used for personality description. The data shows that Goldberg's trait descriptors vary in how they are used for women versus men. Below are the frequency distributions for the personality descriptors. Table 3 showcases the frequency of co-occurrences of traits, and Figure 5 shows a word cloud showing personality descriptors that were found to be shared between men and women in the text.

The word cloud indicates the frequency of descriptors that were found to be shared about men and women. "Kind" is by far the most used descriptor. This is followed by *"Cold"*, then *"Happy", "Quiet",* and subsequently *"Nervous", "Relaxed", "Neat",* and *"Sad"*. These personality descriptors can also be seen in teams of emotional states that people could be in when they are referred to in the text. At a lower frequency, words like *"Anxious", "Tense", "Generous", "Shy",* and *"Rude"* are used to refer to people in the books and such descriptors are often used in referring to traits that have been observed over time.



Figure 5: Personality Descriptors Common to Men and Women

Figure 5 presents common descriptive traits between men and women that can be categorised under Extraversion, including *"Energetic"*, *"Quiet"*, *"Shy"*, *"Talkative"*, and *"Timid"*. Under Agreeableness, it was *"Cold"*, *"Cooperative"*, *"Generous"*, *"Kind"*, *"Pleasant"*, *"Rude"*, *"Sympathetic"*, and *"Unkind"*. Emotional Stability contained *"Anxious"*, *"Depressed"*, *"Irritable"*, *"Nervous"*, *"Relaxed"*, *"Resentful"*, *"Sad"* and *"Tense"*. Conscientiousness had these traits: *"Careless"*, *"Disorganised"*, *"Efficient"*, *"Neat"*, *"Organised"*, and *"Thorough"*. Lastly, Openness to Experience had *"Creative"*, *"Imaginative"*, *"Intellectual"*, and *"Philosophical"*. Table 3 shows the gender trait frequency found under the co-occurrence of pronouns and traits. Though these listed traits are common in both men and women, there is an indication from Table 3 that the frequency is higher for men compared to women in most of the shared traits. This means these traits were used more to refer to men than women.

Trait	Male	Female
Extraversion		
bashful	1	
energetic	2	1
extraverted	1	
inhibited		1
introverted		
quiet	103	66
shy	11	20
talkative	3	1
timid	3	4
untalkative		
Agreeableness		
cold	140	114
cooperative	6	2
generous	15	10
harsh		
kind	206	137
pleasant	11	9
rude	20	10
sympathetic	3	3
unkind	1	1
unsympathetic	1	
Emotional Stability		
anxious	22	16
depressed	7	5
irritable	5	1
nervous	34	17
relaxed	32	13
resentful	1	1
sad	30	25
tense	16	9
touchy		3
unenvious		
careless	13	1
disorganised	1	1
efficient	1	3
impractical		

Conscientiousness		
inefficient		1
neat	30	16
organised	6	2
systematic		1
thorough	3	2
unsystematic		
Openness to experience		
creative	10	3
imaginative	2	1
innovative		1
intellectual	10	5
introspective		
philosophical	7	1
uncreative		
unimaginative		
uninformed		
unintellectual		

Table 3: Gender Traits Frequency

The frequency table above (Table 3) presents the frequency of co-occurrences of the Goldberg terms with male and female pronouns. The table showcases the frequency of traits that are used to describe men and women when correlating pronouns that are used alongside the Goldberg traits.



Figure 6: Personality Descriptors for Men

Figure 6 represents an adjective word cloud that shows the most frequently used words to describe men from Goldberg's list of personality descriptors. At a higher level, it follows the sequence of the common words found for both men and women, "Kind" being the most used descriptor. This is followed by *"Cold"*, *"Happy"*, *"Quiet"*, and subsequently *"Nervous"*, *"Extraverted"*, *"Bashful"*, and *"Sad"*. Table 3 expands by showing personality descriptors that were not found for men were *"inhibited"*, *"Introverted"*, *"Untalkative"*, *"Harsh"*, *"Touchy"*, *"Unenvious"*, *"Impractical"*, *"Inefficient"*, *"Systematic"*, *"Unsystematic"*, *"Innovative"*, *"Introspective"*, *"Uncreative"*, *"Unimaginative"*, *"Uninformed"*, and *"Unintellectual"*. This indicates that these are personality traits present in Goldberg's Trait list; however, they are

not frequently used to describe men in this literary text.



Figure 7: Personality Descriptors for Women

Figure 7 clearly shows that women share the majority of their personality trait descriptors with men, such as being referred to as *"Kind"*, *"Cold"*, *"Happy"*, *"Quiet"*, as well as *"Nervous"*, *"Inhibited"*, *"Systematic*" and *"Sad"*. However, Table 3 outlines that words such as *"Bashful"*, *"Extraverted"*, *"Introverted"*, *"Untalkative"*, *"Harsh"*, *"Unsympathetic"*, *Unenvious*, *"Impractical"*, *"Unsystematic"*, *"Introspective"*, *"Uncreative"*, *"Unimaginative"*, *"Uninformed"*, and *"Unintellectual"* were not used when describing women. This indicates that women are not often described in such a manner, and other descriptors are often used to describe such behavioural attributes.

4.4 Topic Modelling – General Corpus

4.4.1 Unsupervised Method – LDA

The analysis of the South African textual data was done using the unsupervised method of the Latent Dirichlet Allocation. This technique involves randomly assigning words to topics that they thematically align to (Thielmann et al., 2020). With the textual analysis of the books, the topics were built per probabilistically chosen themes (The usage and occurrence of words under a singular clustering/theme that have been statistically selected). Figure 8 presents the

top 10 topics yielded upon running the LDA. The selection of the number of topics is founded on something other than a theoretical backing, as Gan and Qi (2021) mention that no research supports the number of topics that need to be selected. The understanding of topic selection is on the premise that if one selects fewer topics, the word distribution becomes narrow. In contrast, the selection of many topics results in a broader selection of words that do not give much depth of analysis (Gan & Qi, 2021). This results in the ten probabilistic topics being selected to observe the distribution of the words.



Figure 8. Topic Distribution for Unsupervised Topic Modelling of SA books.

The above figure (Figure 8) illustrates the topics that were probabilistically selected from the list of South African books through the process of Latent Dirichlet allocation. The selected topics are 10, with each topic summarised by 10 terms randomly selected under each topic. Since the method of textual selection is unsupervised, the model contains words that are selected randomly and create noise for the analysis. In evaluating the model, some terms occur in more than one topic. These are words like *"said"*, *"man"*, *"men"*, *"one"*, and others. The model also picked up some of the Afrikaans words (Topic 7) that were contained in some of the textual data.

Though the model yielded results on the topics contained in the textual data, the topics are unintelligible and provide no insight into the data. The novels that were used are of different genres and contain different themes. That, however, was not picked up by the model. Instead, it produced words that are most used in novels in referring to characters, such as *"said,"* speaking to what a character spoke about; *"man"*, a reference to the gender of the character; and *"one"*, a numerical and individual reference. The topics contain a tiny selection of words, considering the size of the text. This means that more words that better link to personality understanding could be under each topic. Thus, using such a model in unfiltered data does not adequately assist in discerning the personality traits contained in the South African literary text. This means using semi-supervised topic models that include guidance from dictionaries could be a more appropriate method.

4.4.2 Semi-supervised Method

4.4.2.1 Dictionaries

A topic comparison was done on the textual data against an externally compiled dictionary of personality descriptors. Three dictionaries were compiled to be used for comparison purposes. The first dictionary word list contained words unique to the South African Personality Inventory, the Chinese Personality Inventory, and the English Personality descriptors (IPIP list). Some of the words can be found in either of the inventories. However, the way the list was compiled was to start with the SAPI list and then include words that could not be found in the SAPI list, but that were in the English personality descriptors word list. Lastly, the words that could not be found in the SAPI or English word lists but were categorised under the Chinese Personality Inventory were also included. The second dictionary combines the first one, which includes the SAPI list, the English word list and the Chinese personality word list.

The third list contains words informed by Goldberg's adjective list, forming part of the five-factor categorisation (Goldberg et al., 2006). These word lists are split into two: one list with the trait descriptors categorised according to the five factors from the five-factor model, namely, (Extraversion, Openness to experience, Conscientiousness, Neuroticism, and Agreeableness). The second wordlist contained the Goldberg personality list without the categorisation. The division of these dictionary lists into categorised and uncategorised is done to evaluate the distribution of personality traits within each categorisation and outside of it. This is meant to discern the range of representation of such personality traits in South African literary texts.

4.4.2.2 Word comparison

(I) Personality Inventories

The dictionaries that were created were compared against the textual documents. This was done to see the prevalence and representation of the personality traits contained in these dictionaries. Below is a representation of topics following the dictionary categorisation. The first categorisation would be the first wordlist comparing the categorised dictionary of personality traits from the three inventories (SAPI, CPAI and English) against the textual data.

a. Three personality inventory dictionary

Number of terms	SAPI	English	Chinese
[1,]	kind	different	suspicious
[2,]	serious	nice	modest
[3,]	proud	strange	average
[4,]	fair	aware	honourable
[5,]	honest	private	impose
[6,]	determined	prepared	supportive
[7,]	grateful	careful	unworthy
[8,]	patient	ordinary	boastful
[9,]	understanding	curious	thrifty
[10,]	responsible	nervous	hij
[11,]	cruel	brave	ich
[12,]	loving	willing	zijn
[13,]	friendly	bitter	der
[14,]	traditional	lonely	und
[15,]	useless	decent	niet
[16,]	pretending	anxious	voor
[17,]	intelligent	impatient	naar
[18,]	generous	extraordinary	dan

[19,]	pleasant	persuade	als
[20,]	content	faithful	net
[21,]	shy	coward	moet
[22,]	jealous	joking	domingo
[23,]	reserved	vicious	auf
[24,]	stubborn	accepting	weet
[25,]	emotional	offended	den
[26,]	fearful	spiritual	nog
[27,]	humble	frustrated	hem
[28,]	sensitive	reliable	wil
[29,]	critical	enthusiastic	das
[30,]	demanding	dependent	zegt
[31,]	organised	energetic	hoe
[32,]	complaining	believer	ein
[33,]	peaceful	accomplish	war
[34,]	caring	creative	heeft
[35,]	helpful	predictable	sie
[36,]	respectable	grumpy	tot
[37,]	arrogant	apprehensive	twee
[38,]	loyal	stylish	oor
[39,]	strict	seductive	deur
[40,]	independent	pretentious	ook
[41,]	noisy	irresponsible	hier
[42,]	logical	dishonest	wie
[43,]	protective	manipulative	mit
[44,]	dedicated	expressive	want
[45,]	musical	vindictive	mijn
[46,]	cheerful	sadistic	mense
[47,]	selfish	deceptive	meer
[48,]	irritating	accountable	nou
[49,]	respectful	analytical	dem
[50,]	reckless	pessimistic	waar
[51,]	passionate	adaptable	weer
[52,]	encouraging	submissive	kyk
[53,]	thorough	empathetic	mir
[54,]	greedy	perfectionist	heerden
[55,]	obedient	untrusting	nicht
[56,]	mature	principled	von
[57,]	welcoming	said	man
[58,]	creative	back	alles
[59,]	aggressive	one	sich
[60,]	integrity	just	maak
[61,]	cooperative	like	staan
[62,]	guiding	know	dis
[63,]	competent	now	wir
[64,]	spontaneous	time	iets
[65,]	meticulous	get	mich
[66,]	tensed	man	geen

[67,]	straightforward	see	heb
[68,]	troublesome	looked	groot
[69,]	courageous	two	laat
[70,]	disciplined	head	jaar
[71,]	coping	can	daardie
[72,]	secretive	going	weg
[73,]	inquisitive	got	jij
[74,]	compassionate	eyes	soos
[75,]	dreamer	still	aus
[76,]	artistic	around	bij
[77,]	trustworthy	way	hebben
[78,]	playful	door	hand
[79,]	talented	right	dink
[80,]	forgiving	hand	zich
[81,]	consistent	asked	eerste
[82,]	gossiping	away	doen
[83,]	shamed	something	eine
[84,]	truthful	thought	omdat
[85,]	investigative	think	sit
[86,]	interfering	face	hoop
[87,]	considerate	come	waren
[88,]	articulate	want	musa
[89,]	adventurous	white	toen
[90,]	influential	look	loop
[91,]	competitive	knew	noch
[92,]	optimistic	made	nee
[93,]	humane	even	ander
[94,]	talkative	little	alleen
[95,]	appreciative	black	ist
[96,]	purposeful	left	teen
[97,]	intimidating	took	kry
[98,]	rebellious	room	hatte
[99,]	agreeable	first	door
[100,]	advising	put	aber
[101,]	provoking	say	hoor
[102,]	knowledgeable	tell	terug
[103,]	exaggerate	turned	almal
[104,]	accommodating	people	huis
[105,]	exploiting	take	hou
[106,]	undermining	went	wanneer
[107,]	dutiful	long	zal
[108,]	authoritative	came	wees
[109,]	needy	told	dass
[110,]	perceptive	good	goed
[111,]	impulsive	make	sagte
[112,]	progressive	never	amanzi
[113,]	neurotic	behind	zou
[114,]	outspoken	old	einen

[115,]	delinquent	well	für
[116,]	sociable	car	begin
[117,]	vivacious	hands	vra
[118,]	assertive	felt	iemand
[119,]	captivating	peekay	zij
[120,]	punctual	last	mein
[121,]	constructive	big	onder
[122,]	uplifting	night	keer
[123,]	enterprising	front	stem
[124,]	abusive	saw	lang
[125,]	denigrating	another	kon
[126,]	materialistic	voice	ben
[127,]	approachable	much	tussen
[128,]	visionary	wanted	drie
[129,]	tolerant	nothing	skiet
[130,]	likeable	place	bei
[131,]	individualism	yes	nico
[132,]	flexible	day	word
[133,]	humorous	side	nach
[134,]	didactic	must	nooit
[135,]	temperamental	open	hennie
[136,]	wrathful	jade	kijkt
[137,]	tidiness	small	agter
[138,]	depressive	police	daai
[139,]	argumentative	house	einem
[140,]	disciplining	thing	paar
[141,]	religiosity	stood	langs
[142,]	peacekeeping	work	niks
[143,]	one	let	werk
[144,]	like	though	moeder
[145,]	says	walked	dood
[146,]	can	always	okkie
[147,]	man	phone	hierdie
[148,]	said	next	staat
[149,]	people	looking	meine
[150,]	know	heard	later

Table 4. Categorised Wordlist Dictionary

The above table (Table 4) presents the word comparisons of the categorised dictionary containing words from the SAPI inventory, the CPAI inventory and the English Inventory against the literary textual data. The results of Table 5 indicate that words are found in both inventories and textual data.

For the first topical list (SAPI), the table shows the lists of terms that are both the same and correlate to the ones found in the SAPI. Terms number 1 to 142 are personality descriptors that can be located on both the textual data and the SAPI-formulated dictionary. In the

original wordlist containing the personality descriptors from the SAPI, the total number of words presented under this category is 186, while the topic managed to correspond to 142 words. The number of terms that were not found to correspond with the textual data was 44 words.

The second topical list (English) contains words classified as personality descriptors in English. The table results indicate that in addition to the SAPI words found under the English descriptors, some of the words found exclusively in English correspond with those found in the textual data. From term 1 to term 56 are words extracted from the textual data corresponding to English personality descriptors. The dictionary wordlist containing the original English descriptors contained 68 words. This means that there are about 13 words that were not found in the textual data falling under the English descriptors.

The third topical list (Chinese) contains words unique to the CPAI inventory. Though the list appears short, some words were not repeated as they were found under the English dictionary list or the SAPI dictionary list. The results present term 1 to term 9 as words that correspond with the dictionary from the textual data. The number of words from the original dictionary contained 10 personality traits. The list of corresponding traits was 9 words. This means that most of the personality traits found to be unique in the Chinese dictionary correspond to the textual data.

The method of textual extraction was semi-supervised due to the assistance of dictionaries. This is clearly illustrated by how the model still manages to present words categorised under each topic but does not indicate any thematic relatedness. The model picking up these words is due to the functionality of the LDA, which finds patterns of words that repeat together and frequently occur (Péladeau & Davoodi, 2018). Li et al. (2018) mention that the model's effectiveness can be lowered for massive short texts due to the extraction of noise (thematically unrelated words) and sparsity. As such, for future research, there needs to be a consideration of using techniques such as Topic-Noise Discriminator (TND), Noiseless LDA (NLDA), and Guided Topic-Noise Model (GTM) for topic modelling (Li et al., 2018).

The table below presents the list of terms from each topic not found in the textual data. The results illustrate personality descriptors that are presented as phrases. Phrases such as *"solving problems for others"* or *"calm yourself"* were not found. This could be because the analysis model searched for singular words and not phrases as personality descriptors. Furthermore, these personality traits might not have been found in the textual data because they are not

commonly used as personality descriptors within the South African literary language. In the SAPI categorisation, words such as *"communicative"*, *"storyteller"*, *"heedful"*, *"timeous"*, *"perseverant"*, *"discriminative"*, and *"appeasing"* were not found when the dictionary comparison was done. These personality descriptors are found under the South African Personality Inventory (SAPI); however, they are not used when describing characters in literary texts. These words need to be clarified in the context of personality descriptions. In the English categorisation, words such as *"workaholic"* and *"observative"* were not found when the dictionary comparison was performed. This indicates that these words are not commonly used words when describing individuals. Furthermore, the model was not programmed to pick up bigrams (two words), which could also limit the extraction of any words that could have been used, such as *"physical appearance"*.

SAPI	English	Chinese
emotional sharing	mean-spirited	peace-keeper
pleasure sharing	workaholic	
pleasure seeking	calm yourself	
communicative	non-believer	
extrovert/introvert	observative	
storyteller	physical appearance	
self-centred	easily distracted	
verbally aggressive	good listener	
satisfying others	being able to follow	
community involvement	narrow-minded	
heedful	detail-oriented	
solving problems for others		
career-oriented		
hard-working		
performance-oriented		
timeous		
future-oriented		
perseverant		
follow-up		
absent-minded		
attention-seeking		
self-confidence		
self-respectful		
obsessive/compulsive		
balancing life		
even-tempered		
short-tempered		
concrete work		

self-insight		
social intelligent		
academically oriented		
eager to learn		
open-minded		
prim and proper		
fashion conscious		
like to travel		
discriminative		
open for others		
appeasing		
good relations with another		
well-mannered		
role model		
aspirations for others		
thought-provoking		

Table 5. Absent Personality descriptors

b. Combined personality inventory dictionary

The second wordlist comparison is done on the combined list of words dictionary containing all the words from the SAPI, CPAI and English without categorisation. Table 6 below shows the results of the comparison.

Number of terms	Terms
[1,]	kind
[2,]	different
[3,]	nice
[4,]	strange
[5,]	serious
[6,]	aware
[7,]	private
[8,]	prepared
[9,]	careful
[10,]	proud
[11,]	fair
[12,]	ordinary
[13,]	honest
[14,]	curious
[15,]	determined
[16,]	nervous
[17,]	grateful
[18,]	brave
[19,]	willing

[20,]	patient
[21,]	bitter
[22,]	lonely
[23,]	understanding
[24,]	decent
[25,]	responsible
[26,]	cruel
[27,]	loving
[28,]	anxious
[29,]	friendly
[30,]	traditional
[31,]	useless
[32,]	pretending
[33,]	intelligent
[34,]	impatient
[35,]	generous
[36,]	pleasant
[37,]	content
[38,]	extraordinary
[39,]	suspicious
[40,]	shy
[41,]	jealous
[42,]	reserved
[43,]	stubborn
[44,]	emotional
[45,]	fearful
[46,]	humble
[47,]	sensitive
[48,]	critical
[49,]	persuade
[50,]	demanding
[51,]	faithful
[52,]	organised
[53,]	modest
[54,]	complaining
[55,]	peaceful
[56,]	average
[57,]	caring
[58,]	arrogant
[59,]	helpful
[60,]	respectable
[61,]	coward
[62,]	loyal
[63,]	strict
[64,]	independent
[65,]	noisy
[66,]	logical
[67,]	joking

[68,]	creative
[69,]	protective
[70,]	dedicated
[71,]	musical
[72,]	cheerful
[73,]	vicious
[74,]	accepting
[75,]	selfish
[76,]	offended
[77,]	irritating
[78,]	respectful
[79,]	spiritual
[80,]	reckless
[81,]	frustrated
[82,]	reliable
[83,]	passionate
[84,]	encouraging
[85,]	thorough
[86,]	honourable
[87,]	greedy
[88,]	obedient
[89,]	mature
[90,]	welcoming
[91,]	aggressive
[92,]	enthusiastic
[93,]	integrity
[94,]	cooperative
[95,]	guiding
[96,]	competent
[97,]	dependent
[98,]	energetic
[99,]	spontaneous
[100,]	believer
[101,]	meticulous
[102,]	tensed
[103,]	accomplish
[104,]	straightforward
[105,]	troublesome
[106,]	courageous
[107,]	disciplined
[108,]	coping
[109,]	secretive
[110,]	inquisitive
[111,]	compassionate
[112,]	impose
[113,]	dreamer
[114,]	artistic
[115,]	supportive

[116,]	predictable
[117,]	trustworthy
[118,]	playful
[119,]	talented
[120,]	forgiving
[121,]	grumpy
[122,]	apprehensive
[123,]	consistent
[124,]	gossiping
[125,]	unworthy
[126,]	shamed
[127,]	truthful
[128,]	investigative
[129,]	interfering
[130,]	considerate
[131,]	articulate
[132,]	stylish
[133,]	seductive
[134,]	adventurous
[135,]	influential
[136,]	pretentious
[137,]	competitive
[138,]	optimistic
[139,]	humane
[140,]	talkative
[141,]	appreciative
[142,]	purposeful
[143,]	intimidating
[144,]	rebellious
[145,]	agreeable
[146,]	irresponsible
[147,]	dishonest
[148,]	advising
[149,]	manipulative
[150,]	provoking
[151,]	knowledgeable
[152,]	exaggerate
[153,]	accommodating
[154,]	expressive
[155,]	exploiting
[156,]	undermining
[157,]	vindictive
[158,]	dutiful
[159,]	authoritative
[160,]	needy
[161,]	perceptive
[162,]	impulsive
[163,]	progressive

[164,]	neurotic
[165,]	outspoken
[166,]	sadistic
[167,]	deceptive
[168,]	delinquent
[169,]	sociable
[170,]	vivacious
[171,]	boastful
[172,]	assertive
[173,]	captivating
[174,]	punctual
[175,]	constructive
[176,]	uplifting
[177,]	enterprising
[178,]	accountable
[179,]	abusive
[180,]	denigrating
[181,]	materialistic
[182,]	analytical
[183,]	pessimistic
[184,]	adaptable
[185,]	approachable
[186,]	visionary
[187,]	tolerant
[188,]	likeable
[189,]	individualism
[190,]	flexible
[191,]	humorous
[192,]	submissive
[193,]	didactic
[194,]	empathetic
[195,]	temperamental
[196,]	wrathful
[197,]	tidiness
[198,]	depressive
[199,]	argumentative
[200,]	perfectionist
[201,]	thrifty
[202,]	disciplining
[203,]	untrusting
[204,]	religiosity
[205,]	principled
[206,]	peacekeeping
[207,]	said
[208,]	one
[209,]	like
[210,]	back
[211,]	know

[212,]	just
[213,]	man
[214,]	now
[215,]	can
[216,]	time
[217,]	see
[218,]	get
[219,]	two
[220,]	people
[221,]	even
[222,]	come
[223,]	says
[224,]	want
[225,]	say
[226,]	going
[227,]	way
[228,]	away
[229,]	never

Table 6. Combined wordlist

The above table (Table 6) presents terms classified in the personality descriptor combined list. The original dictionary contains 264 words representing personality descriptors. The comparative words that are present in this list are 206. This means that about 57 words cannot be found in the textual data. Since the dictionary contains the exact words as the previously categorised ones, the different words will be the same, and there will be differences in the words.

(II) Goldberg Personality Inventories

c. Goldberg Personality Inventory dictionary

The Goldberg personality item tool (Goldberg et al., 2006) contains words that form part of one of the most used personality classifiers, the Big Five model. The words that Goldberg presents in the item tool were used to formulate a dictionary. This dictionary was used to extract these terms from literary textual data. The dictionaries contain words that have been categorised according to the five factors, and one has the words combined. The table below shows the results of the extraction.

Number of terms	Extraversion	Agreeableness	Emotional Stability	Conscientiousness	Openness
[1,]	quite	kind	sad	neat	intellectual
[2,]	shy	cold	nervous	organised	creative
[3,]	timid	harsh	anxious	careless	philosophical
[4,]	energetic	generous	relaxed	thorough	imaginative
[5,]	talkative	pleasant	tense	efficient	innovative
[6,]	bashful	rude	depressed	impractical	uninformed
[7,]	introverted	sympathetic	irritable	systematic	unimaginative
[8,]	inhibited	cooperative	resentful	inefficient	introspective
[9,]	extraverted	unkind	touchy	disorganised	uncreative
[10,]	ich	unsympathetic	said	unsystematic	unintellectual
[11,]	der	like	one	one	hij
[12,]	und	says	back	said	zijn
[13,]	kommandant	can	just	man	niet
[14,]	kramer	one	know	men	voor
[15,]	auf	just	like	like	naar
[16,]	den	back	now	people	dan
[17,]	das	know	time	two	net
[18,]	von	see	man	must	moet
[19,]	ein	time	looked	now	domingo
[20,]	war	say	get	back	weet

Table 7. Five-Factor Word Comparison

Table 7 presents the results of the word comparison between the five-factor classified dictionary and the literary textual data. The model extracted words that can both be found in the dictionary and the literary textual data. The dictionary categorisation was classified under (Extraversion, Agreeableness, Emotional stability, Conscientiousness, and Openness) and each categorisation had 10 terms. The extraversion and emotional stability categorisation consist of 9 words the model extracts corresponding with textual data. For the extraversion categorisation, the word that did not correspond to the textual data was *"untalkative"*, and with the emotional stability categorisation, the word that did not correspond to the textual data was *"unenvious"*. The model extracted some words but did not fall under the dictionaries that were random and, as such, did not form any analytical relevance. These can be noted as the noise extracted by the model, which was not controlled for. This, however, does not mean the model cannot classify other terms under each topic. The current extraction showcases corresponding words between the textual and the dictionary.

4.5 Topic Modelling – Adjective distribution

4.5.1 Unsupervised LDA

It has been established that adjectives are often used to express personality traits. An investigation into cluster formation was done using a topic model on a dataset of adjectives from the textual data. The adjectives used in the textual data were isolated and captured in a dataset to be clustered in a probabilistic manner. The process of extracting the adjectives involved using the *UDpipe* package (Wijffels et al, 2020), which isolated words corresponding to the part of speech of adjectives. This yielded only words that were classified as adjectives and were captured in a dataset. The overall file contained words that were not in English, thus requiring further cleaning and removal of the Afrikaans and Dutch-related words.

The LDA was followed as a topic modelling of choice as it allows for the capturing of thematic clustering of topics, which can be compared against existing personality inventories. Ten topics were created for this analysis to compare against the SAPI inventory. The SAPI inventory contains 9 clusters. As such, ten topics were created to compare the extraction of the 9 topics against the 9 clusters and to check if the 10th topic yields or can yield a different clustering. This comparative analysis aimed to discover potential correlations and discrepancies between the personality trait clustering derived from traditional inventories and the alternative approach offered by topic modelling. These results are reflected in Figure 9 below.

The extraction of words through LDA is subject to randomness. Hence, a seed is used in the coding to ensure the reproducibility of the same set of words. Though the word extraction might be random, the words under each topic are not random, as the Dirichlet functionality guides them. This looks at the pattern of repeated words occurring frequently, and these words are similar to each other (Péladeau & Davoodi, 2018). Figure 9 below contains ten charts of 10 topics with 10 bars containing words that were probabilistically extracted under each topic. As noted in Figure 8, the topics often do not have thematic labelling as that is usually connotated by the researcher based on the observable relation between words.

The thematic investigation was done based on the researcher's interpretation of the association of the words per topic. The 10 topics showcase the clustering of related adjectives that frequently occur together. This classification shows basic descriptors used in any given context. With the topic modelling approach being an approach that focuses on a bag of words negating any contextual relevancy, it cannot be said what the descriptions are. However,

because these words were extracted from a literary text, each topic represents a descriptive story. To formulate that, a larger term selection under each topic needs to be done. The topics below do not contain all the words, only ten terms. The terms selected under each topic are like those found in Figure 4, which looks into SA literature's top 20 frequently used adjectives. The distribution of these terms in these topics shows themes of colour, time, value, size, and age.



Figure 9: Adjective Topic Modelling

4.5.2 Supervised LDA

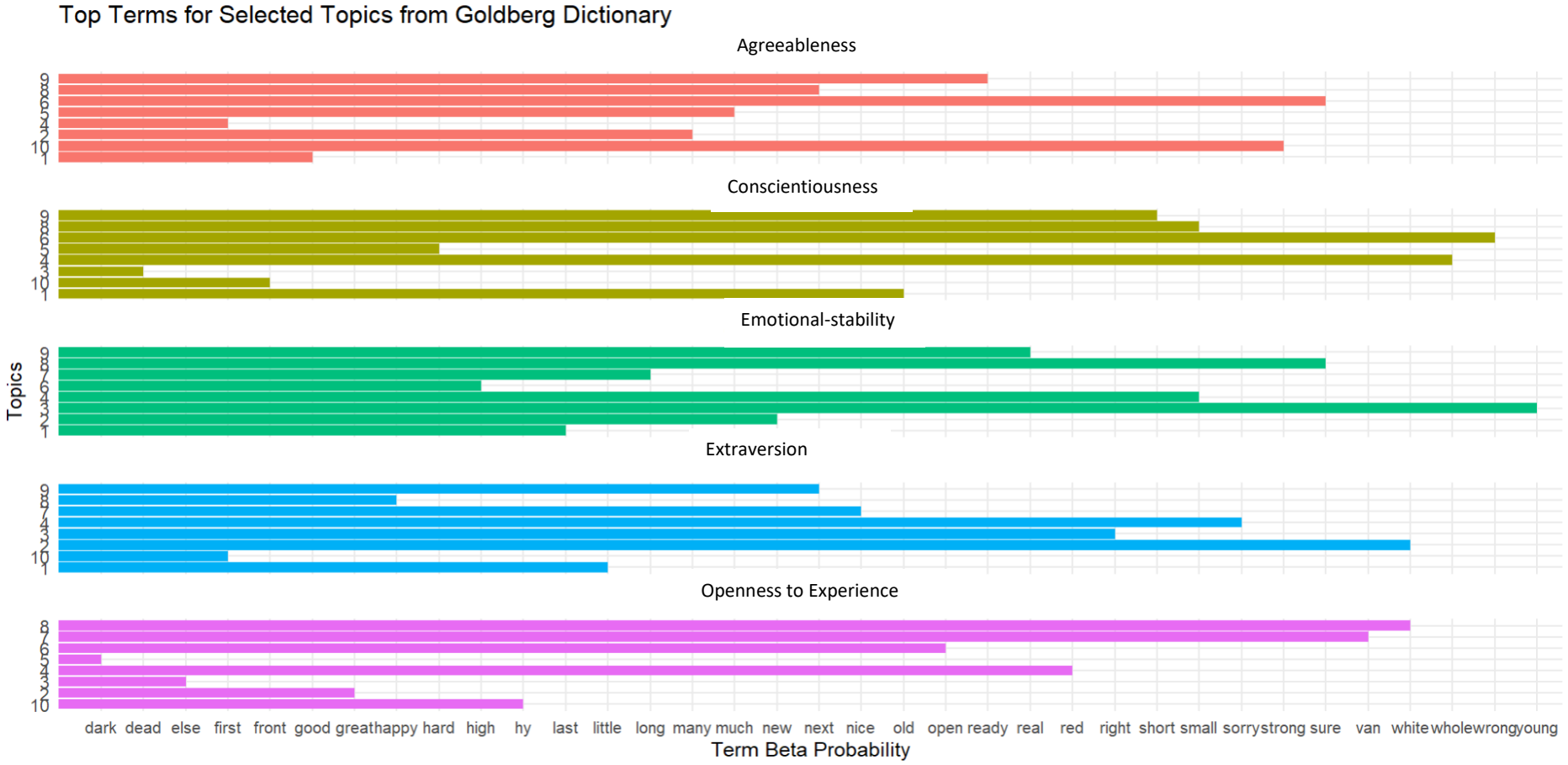


Figure 10: Supervised Adjective Classification

A guided approach to understanding personality distribution was implemented using Goldberg's list as the dictionary in the topic modelling. The approach implemented in Figure 10 topics showcases the frequency of terms from each of the topics under the five-factors model. The topic modelling was done in the adjectives-only corpus. The above figure (Figure 10) takes the ten topics that were extracted from the unsupervised model and determines which term is frequent under agreeableness (sure), Conscientiousness (wrong), emotional stability (young), as well as extraversion and openness (white), using the beta probability. This measures the probability of term occurrence in each topic.

In selecting 8 out of the ten topics, it can be noted that the same topics were not selected under each personality classification. This means different terms were frequent from the different topics under each personality classification. Topic 4 and 8 can be seen occurring in all the classifications. This shows that there are traits that can be attributed to each classification of personality; furthermore, it leaves room for further investigation. The openness classification is the only one with no topics 1 and 9, which could speak to the kind of words understood as forming part of the personality construct of openness to experience.

4.6 Conclusion

The results show the dissection of the South African literary text from text corpus to tokenised words used for analysis. The text was assessed for readability to gauge the standard of descriptive understanding of the words, and it was found that much of the population could understand the words. A distribution of adjectives was measured to ascertain the most frequent adjectives in the text. This led to the establishment of co-occurrence of adjectives with a selected list of pronouns. This helped in discerning any personality differences between men and women. Lastly, the topic modelling technique was explored both in an unsupervised manner (overall text and adjectives) and a supervised manner (dictionaries). More investigations are needed to use the various text mining techniques explored in these results and those not used.

Chapter 5 – Discussion

5.1 Introduction

The chapter discusses the results from Chapter 4. This is done through interpreting and contextualising the results within the presented literature and the research question that was investigated. The contextualisation of the results was discussed concerning the implication of using text mining to understand personality traits in South Africa.

5.2 Research Question 1 – How are personality descriptors classified using text mining?

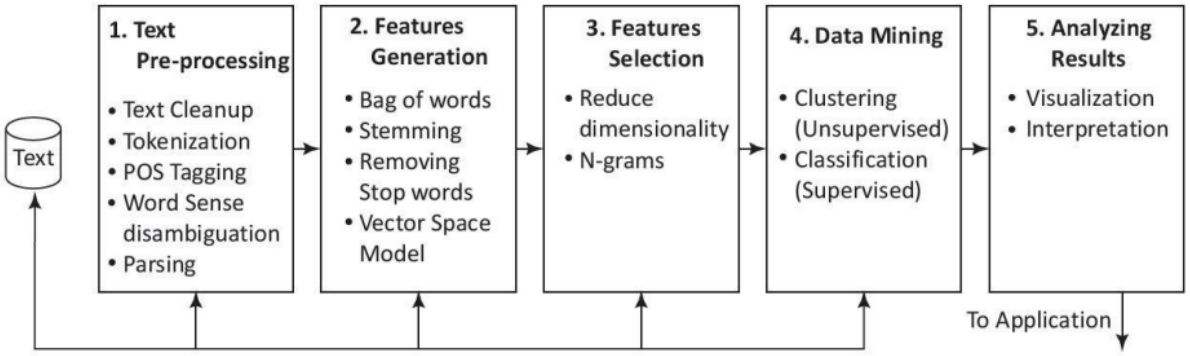


Figure 11: Text Mining Framework. Source: (Kamal & Saxena, 2019)

5.2.1 Corpus analysis

The application of text mining in classifying personality descriptors was a different approach for analysing the descriptive nature of intricate dimensions of character personalities in South African literature. This allowed for the enriching qualitative observations drawn from the corpus and matrices, offering both breadth and depth to the analysis. The textual data used for this study were 60 South African fictional literary texts (see appendix) selected from the library based on their format to fit in with the analysis. The sample of textual data was suitable because each novel contains a minimum of 200 pages, meaning that the level of textual preprocessing and analysis was extensive.

In starting with using text mining in psychology, the main objective is to understand the use of adjectives to describe an individual's behaviour and state of being. Preliminary data preprocessing was done to ensure that only the parts of speech relating to descriptors were left. A process referred to as tokenisation allowed for the singular extraction of texts. This means each word is referred to by its singularity instead of its place in a sentence. A large corpus of the text was therefore formed containing all the words per book, which resulted in a large sum of text to process and go through with no pre-established thematic alignment.

5.2.2 Textual readability

An aspect of textual readability was established to ensure that each text met a standard of comprehensibility by anyone and could easily be used in a daily conversation. Furthermore, personality descriptors must be words people can easily use to describe people and their behaviours. This was notably important in formulating the Chinese Personality Assessment Inventory (CPAI). Using Chinese literary text, a level of readability and comprehension had to be established before any trait and cluster formulation (Cheung et al., 1996b, 2013). The Flesch index became essential in quantifying the text's readability level. It was found that only one book had an index between 50-60, which was 58.6, meaning that the text can be understood by individuals who have a high school level of reading. This means that descriptors can be communicable and comprehended by the majority of the population.

5.2.3 Part of speech tagging - Adjectives

Part-of-speech tagging (POS) was used to classify the different roles that words occupy in each textual and speech format. This was paramount in extracting adjectives used as forms of descriptors, either the description of objects, state of being or individuals. These are characterisations of Indo-European languages such as English, where there are five taxonomies of adjectives, which include social relations, physical characteristics, description of utility, states of being/conditions and dispositional traits (Grzybowski et al., 2021). This is best illustrated by Figure 4, which shows the frequency of the top 20 adjectives found in the literary text.

A variation of adjectives was shown with those frequently used, often referring to states of being and less so on the reference of humans, which is seen from words like *"Other"*. Words such as *"Young", "Great", "Old",* and *"Good"* can be used to refer to humans and their seen attributes as well as personality traits, which link to more states of being and dispositional traits (Grzybowski et al., 2021). These are seen as being frequently used; however, though they can be attributed to personality traits, there is no direct link that they refer to an individual or an individual's behaviour.

A much more directional approach was followed, allowing for the determination of personality trait usage of adjectives used about people. A list of nouns and pronouns that are used to make mention of individuals was constructed. This was done because, though the POS tagging could have been used, it would not have allowed using both pronouns and the selected nouns list. Further, a list of Goldberg's personality traits was used to locate the use of personality attributes in the textual data through the cooccurrence of both pronouns and

adjectives in each sentence. The universal interpretation of Goldberg's trait list made it easier to be chosen as an anchor for understanding the representation of personality traits in South African literary texts. This allowed for the observation of used and not commonly used adjectives as personality traits and the gender difference in differentiating between personality traits used for men and women.

Certain words appear in the description of people from Goldberg's list, and others do not in the reference. This means that these personality traits are not commonly used in this text. Classifying them according to the five-factor model under extraversion, words like *"Untalkative", "Bashful", "Introverted",* and *"Inhibited"* were not picked up by the search. With agreeableness, it was *"Unsympathetic", "Harsh",* and *"Unkind",* which did not appear when a cooccurrence with pronouns was done. Under emotional stability, it is *"Resentful", "Tense",* and *"Unenvious".* At the same time, conscientiousness is *"Inefficient", "Disorganised", "Systematic", "Unsystematic",* and *"Impractical",* and openness to experience is *"Introspective", "Unimaginative", "Uncreative", "Unintellectual",* and *"Uninformed".* A notable reason behind this could be the cultural and linguistic usage of such descriptors about people.

The personality traits shown in this list appear to be negative expressions compared to their referenced positive/opposite counterpart descriptors. As such, there is an indication that such negative expressions are not present when describing people in the South African context while recognised in Western communities. This can be linked to the difference between individualistic and collectivist cultures. Research by Bhawuk (2017) presents that the concept of self between individualistic and collectivist cultures tends to differ in relation to their communities. Individualistic communities are inclined to self-dependency, while collectivist communities are prone to communal dependence. This often leads to a difference in communication between the two cultures. The expression within collectivist cultures can be viewed as being more high-context, meaning their communication is contextually based (Bhawuk, 2017; Triandis, 2018).

On the other hand, individualistic cultures tend to communicate in a low-context manner, meaning their expression is more content-based (Bhawuk, 2017). This can be seen in how individualists are prone to attribute internal traits as the cause of an outcome rather than external occurrence as being a cause, which is often the collectivists' viewpoint. Terms such as *"Unimaginative", "Uncreative", "Unintellectual",* and *"Uninformed"* can be seen as

descriptors of internal traits as opposed to environmental occurrences. As such, the negative attributes were not found in the parts-of-speech tagging. Therefore, it can be linked to the difference collectivists have in describing individuals versus individualists and minimal reference to such terms in contextual settings.

5.2.4 Part of speech tagging – Gender differences

A similar pattern was observed in Table 3 from the results section when looking at the cooccurrence of pronouns and personality traits from Goldberg's list. Words such as *"Inhibited", "Introverted", "Untalkative", "Harsh", "Touchy", "Unenvious", "Impractical", "Inefficient", "Systematic", "Unsystematic", "Innovative", "Introspective", "Uncreative", "Unimaginative", "Uninformed",* and *"Unintellectual"* are not so popular when referring to men or male figures. On the other hand, words such as *"Bashful", "Extraverted", "Introverted", "Untalkative", "Harsh", "Unsympathetic", "Unenvious", "Impractical", "Unsystematic", "Introspective", "Uncreative", "Unimaginative", "Uninformed",* and *"Unintellectual"* is not used in reference to women or female figures. Research on gender and personality has shown that there are few gender differences when it comes to personality (Laher et al., 2020). This has been observed across numerous collectivist cultures such as Nigeria, India, Botswana, and Ethiopia (McCrae & Terracciano, 2005), Zimbabwe, and South Africa (among black South Africans) (Costa et al., 2001).

The parts-of-speech tagging emphasises this by indicating more common personality traits about gender identities. The personality traits that are not interconnected with males tend to fall under *"Inhibited", "Introverted", "Untalkative"* (Extraversion), *"Harsh"* (Agreeableness), *"Touchy", "Unenvious"* (Emotional Stability), *"Impractical", "Inefficient, "Systematic" "Unsystematic"* (Conscientiousness), *"Innovative", "Introspective", "Uncreative", "Unimaginative", "Uninformed", and "Unintellectual"* (Openness-to-experience). For women, they can be classified as follows, *"Bashful", "Extraverted", "Introverted", "Untalkative"* (Extraversion), *"Harsh", "Unsympathetic"* (Agreeableness), *"Unenvious"* (Emotional Stability), *"Impractical", "Unsystematic"* (Conscientiousness), and *"Introspective", "Uncreative", "Unimaginative", "Uninformed",* and *"Unintellectual"* Openness-to-experience).

Though personality traits among gender identities are often not different, a common variation of difference is found between men and women. Laher et al. (2020) presented research findings from personality researchers who found that the difference exists between men and women on three FFM clusters. These clusters are Neuroticism, Agreeableness and

Extraversion. When the traits under these clusters were examined, women often scored higher in Neuroticism and Agreeableness, while men were higher in Extraversion (Costa et al., 2001; Laher et al., 2020; McCrae & Terracciano, 2005). However, this was different when looking at the cooccurrences between pronouns and personality descriptors used by South African authors in their literary texts.

Table 3 presents the frequency of traits observed between men and women. Some of the traits include *"Cold", "Nervous",* and *"Anxious"* (Neuroticism), as well as *"Sympathetic"* and *"Kind"* (Agreeableness) with *"Talkative" and "Energetic"* (Extraversion). These traits rank high in both men and women regarding the terms used in reference to men and women in SA literary texts. The variation could be a result of the cooccurrence of picking up the existence of pronouns alongside the selected personality traits (Goldberg) with minimal consideration of context. A word such as *"Cold"* was ranked as a high personality trait for men and women when a pronoun-adjective cooccurrence was done.

The exploration per cooccurrence extraction indicates that *"Cold"* could refer to the environment where the individual is located. The model would pick it up as a personality trait due to using an adjective and a pronoun in the same sentence. This is not something that occurred with every trait, but it was common with adjectives that can be used as contextual references. This incorrect association of the model could be the type of text, large corpus, and the model's ability to distinguish between individual and purely contextual references. The evidence indicates that further research is required to understand the differentiation in personality traits and clustering, especially when understanding conscientiousness and openness to experience. The results indicated that aspects of the conscientiousness trait clustering are a common difference between men and women.

5.2.5 Topic modelling

5.2.5.1 Unsupervised modelling

Upon establishing the existing nature of personality traits and the varying uses for individuals, a form of clustering was then tested. The topic modelling approach creates a topical classification of words following themes that link them within and between documents (Gan & Qi, 2021). Topic modelling, being an unsupervised machine learning method, contains different approaches that can be used to cluster words into topics (Gan & Qi, 2021; Thielmann et al., 2020). In this case, the Latent Dirichlet Approach was used because it manages to create a probabilistic thematic grouping of words into topics found in the document (Blei et al., 2003). Therefore, the topics are interpreted based on how they

relate to each other in the text. An overall unguided topic modelling was performed, which involved looking at themed topical grouping of the words from all the texts. This was to establish any linguistic theming of words and capture underlying individual descriptions.

About ten topics were created, and with a larger topic selection, the thematic alignment decreases, leading to a much more increased probability of themes, but that does not yield relatability (Blei et al., 2003). It must be considered that the investigation, in this case, is not looking at textual themes but more at descriptive themes and the use of descriptors. Liu (2016) supports that though unsupervised topic modelling can discover underlying topics, those topics have the potential of not representing what is being investigated. As such, the topics did not yield anything that could be interpretable under the personality description.

An unsupervised topic modelling was tested on an adjective bag-of-words intending to refine the general topic modelling that was done using the corpus data. Using the *UDpipe* package, a function was created to extract all the adjectives in the text into a bag of words. An LDA model was then utilised to group the terms under thematic topics. The presented topics showcase ten themes containing frequently used adjectives, which are then classified according to their relatability in each context. The main objective of the LDA is to investigate the document and then the terms to find thematic relatability (Péladeau & Davoodi, 2018). The main descriptive themes of the topics focus on Colour (*White, Black & Green*), Age (*Old & Young*), Value/Moral value (*Good, Great, Wrong & Right*), Time (*First & Last*), Size (*Big, Small, Heavy & Little*) and Quality (*Old & New*). These themes give a basic understanding that links to no specific personality description but meaning can be attributed based on the knowledge of the South African text. These can be seen as social evaluation traits that look at the generality of collective description such as value, age and colour (as this plays a role in South Africa) (Grzybowski et al., 2021). The approach shows some structure in trait classification; however, it still lacks the link to individual behaviour, which speaks to personality classification.

5.2.5.2 Semi-supervised Modelling

This introduced the element of semi-supervised topic modelling, framed by the personality traits from the three personality inventories (SAPI, English and CPAI) and Goldberg's Five Factor word list. The four personality inventories manage to capture the cultural and universal relatability of descriptors used when characterising an individual's behaviour. The traits from the SAPI were then compared against the overall text to find words corresponding to those from the SAPI and ones that may fall under this classification. Taking into

consideration the overlap that exists between the inventories, words that only appear specifically in each inventory were compared, with the SAPI being the initial base word list. Goldberg's Big Five list stood on its own as its comparability came from the finding of the use of some of the universally recognised personality traits in South African literature. This resulted in a list of words that could be found in the text under each personality inventory. Words that could not be found in each personality inventory were also outlined.

The semi-supervised topic modelling assisted in discerning the type of words that are commonly used and can be understood when communicating behaviour in the text. This allowed us to see the contextual relatability in some of the words used and some not found by the model in the text. With recognisable readability, personality descriptors such as *"communicative", "storyteller", "heedful", "timeous", "perseverant", "discriminative", "appeasing",* as well as *"workaholic", "observative", "untalkative"* and *"unenvious"* were singular words that were not found. This indicates the lack of contextual and cultural relatability of these words as they are not used in literature catering to everyone (Triandis, 2001), Supporting the observation of the cultural reference of personality traits between individualistic cultures versus collectivist cultures (Bhawuk, 2017; Gelfand et al., 2004).

An extension of the semi-supervised method was done on the adjectives-only data. This aimed to determine the traits found in SA literature that are classified under the Five Factors. Goldberg's list was used to guide the section of terms as it is the commonly known personality list and contains fewer descriptors, making the classification much quicker. The model yielded terms associated with those that fell under each topic. These terms were the most frequent in each given topic, which was 10, as a selection of 10 was made. As thematic relatability is the core of topic modelling, the different selection of topics can indicate the topics and terms that were most relatable to the personality classification, giving the occurrence of the descriptors as well.

The use of a method like semi-supervised LDA showed the capabilities of using computerised models in grounding the thematic searches for personality traits. The model was applied to show the presence and use of personality descriptors and filter other personality descriptors that could fall under each inventory. This, however, was unsuccessful as the words included by the model under each inventory did not fall under the classification of a personality descriptor. Textual literature tends not to emphasise personality description

but more on contextual and situational description, leaving less room to adequately get the underlying personality traits (Bhawuk, 2017; Cheung et al., 2008, 2011).

Hence, in undertaking text mining analytics for personality research, the techniques presented in Figure 8 form the basis of textual insight, text preprocessing (tidy text, quanteda), feature generation, feature selection, data mining and analysing results (Kamal & Saxena, 2019). The extension to personality research under data mining can first look at the frequency of descriptors and the most common descriptor used towards individuals, allowing for contextual investigation. Further, looking into gender and personality descriptors using the cooccurrence technique (UDpipe). This investigates the occurrence of selected pronouns and selected personality descriptors in a sentence, allowing for the discernment of the kinds of descriptors used for different gender identities.

Unsupervised topic modelling is a technique that is preferable for text data that holds only adjectives as it looks at the kinds of prevalent thematic descriptions done in the text. Those can be attributed to personality descriptors (LDA). Moreover, opening a way to form new personality clusters based on the natural language usage of descriptors in text. A supervised topic modelling technique also compares existing personality traits and clustering to a different context. As such, gauging the use of descriptors in different contexts and the influence of culture (seeded LDA). It is to be noted that more text mining techniques such as NMF, LSA and Parallel LDA can be used, as well as controlling for things like noisy data. Furthermore, Generative AI techniques can be used alongside text mining techniques to support understanding personality construction in the South African context. This can assist with the analysis of extensive textual data, which was a limitation of the study.

5.3 Research Question 2 – Are there different personality clusters produced using topic modelling?

5.3.1 Topic Modelling – General Corpus

No, topic modelling (LDA) does not produce different personality clusters when employed on literary text data. Topic modelling aims to cluster thematically similar words using an algorithmic model under a singular topic, assuming a document contains multiple topics (Laureate et al., 2023; Péladeau & Davoodi, 2018). This often begins at a document level, down to the word level. In this study, the classification began at the book level, down to the text's words. Figure 8 presents ten topics that were extracted using an unsupervised LDA model. Each topic is identifiable with a number assigned by the model. The topics were

constructed through the frequency of the word occurrence and their repetition (Péladeau & Davoodi, 2018). This yielded a combination of common words in the literature, such as *"says"*, *"said"*, *"back"*, and *"men"*.

The results of the topics show that the model extracted no overt themes. This was because of the large corpus of text with numerous latent themes and no clear guidance concerning the number of topics to select. The number of topics was selected randomly, meaning that some meaningful topics could have been missed. As such, the selected terms under each topic cannot be attributed to personality connotation. Murshed et al. (2023) present research on the various topic-modelling approaches that have been created in an attempt to ensure that topic modelling becomes a robust method of text extraction. They outline that topic models like LDA tend to produce topics that need more semantic understanding due to the noisy data filtered in topics, sparsity and the binary weighting of terms (Murshed et al., 2023). Some researchers dismiss the capacity of the LDA to discern themes in the text, but it helps understand the important words used (Okon et al., 2020). This can be supported by exploring a large corpus of data, such as literary text, with no clear guidance on thematic specificity. The only insight you get comes from word frequency.

5.3.2 Topic Modelling – Adjective Distribution

In understanding that the LDA requires some direction concerning topic extraction and personality classification, the adjectives found in the corpus were extracted as descriptors that often form part of personality traits. Ten topics were selected following the 9 clusters present in the construction of the SAPI, which included a 10th topic to determine any thematic and classification differences. Figure 9 presented these topics and terms; the topics presented followed a similar path as in Figure 8, with the extraction of the frequently used terms under each topic. The exception was the supervised LDA, which used a predefined dictionary containing words from Goldberg's Five Factor model. The visualisation in Figure 10 captures the common topics under each personality classification and the frequent terms under that topic. This can be seen with the classification of Openness-to-experience, as it is the only one that does not have topics 1 and 9 under its occurrence, showing a slight distinction between the term and topic classification.

When investigating the social discourse of same-sex marriage, (Hemmatian et al., 2019) mention that thought topic models can have a limited capacity to speak to constructs such as social cognition; their relevancy lies with statistically aligning in interpretation. This means that topic models such as the LDA cannot be used independently as complex constructs

comprising different themes are challenging to classify. Further, they rely primarily on the social context, which is why they are implementing a more guided approach. As such, constructing an adjectives-only bag of words allowed for the understanding and showed the value of using guided topic models to discover the use and relatability of descriptors. An investigation into large corpora of text and a deeper topical investigation can shed light on the classification of personality using literature and how that translates to the formation of personality descriptors in SA.

5.4 Research Question 3 – What are the common themes found in the South African literary text in comparison to existing personality classifications?

5.4.1 Supervised Text Mining - Frequency

No specific themes were found, but traits that fall under each existing personality cluster were found to be used in the South African literary text. Supervised topic modelling is an approach that utilises pre-established terms to guide the thematic allocation of topics (Blei et al., 2003; Lin et al., 2012; Mcauliffe & Blei, 2007). The method categorises terms that probabilistically align with those introduced into the model. Personality traits under each personality (SAPI, English, CPAI and Goldberg) classification were introduced to extract those that align with each classification.

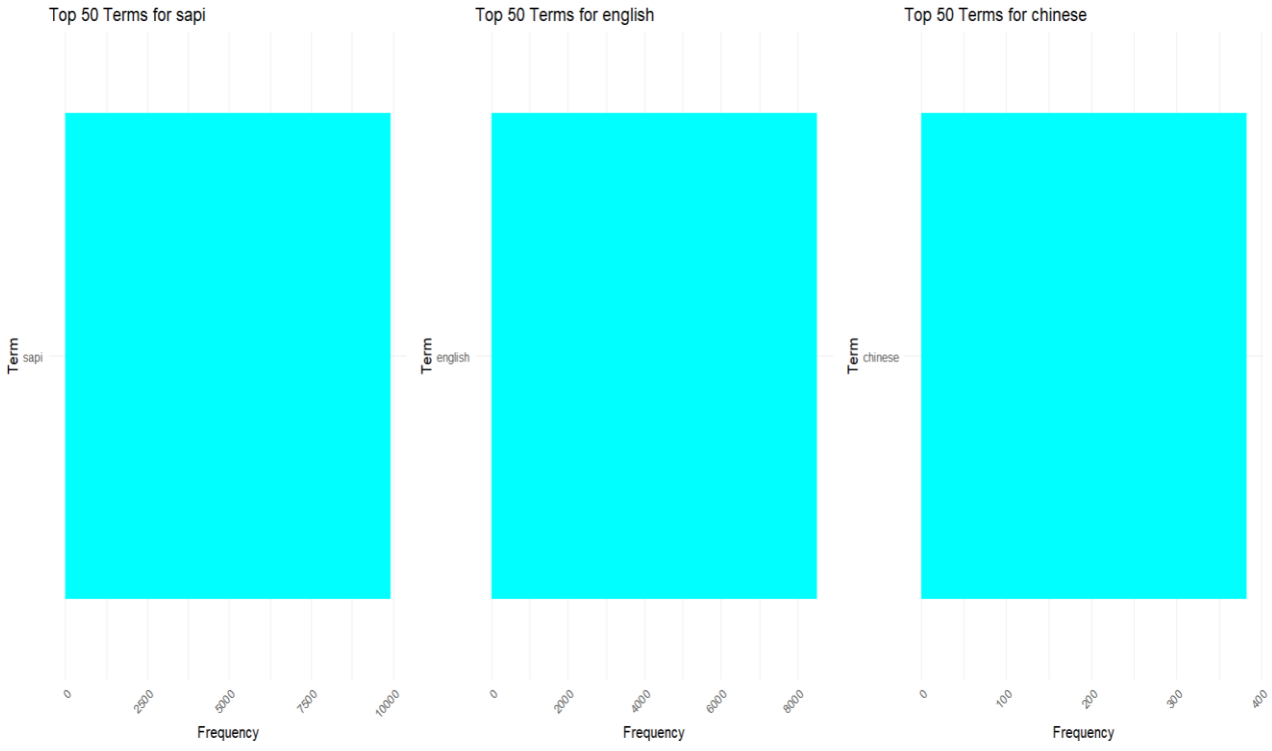


Figure 12: Word List Dictionary Frequency

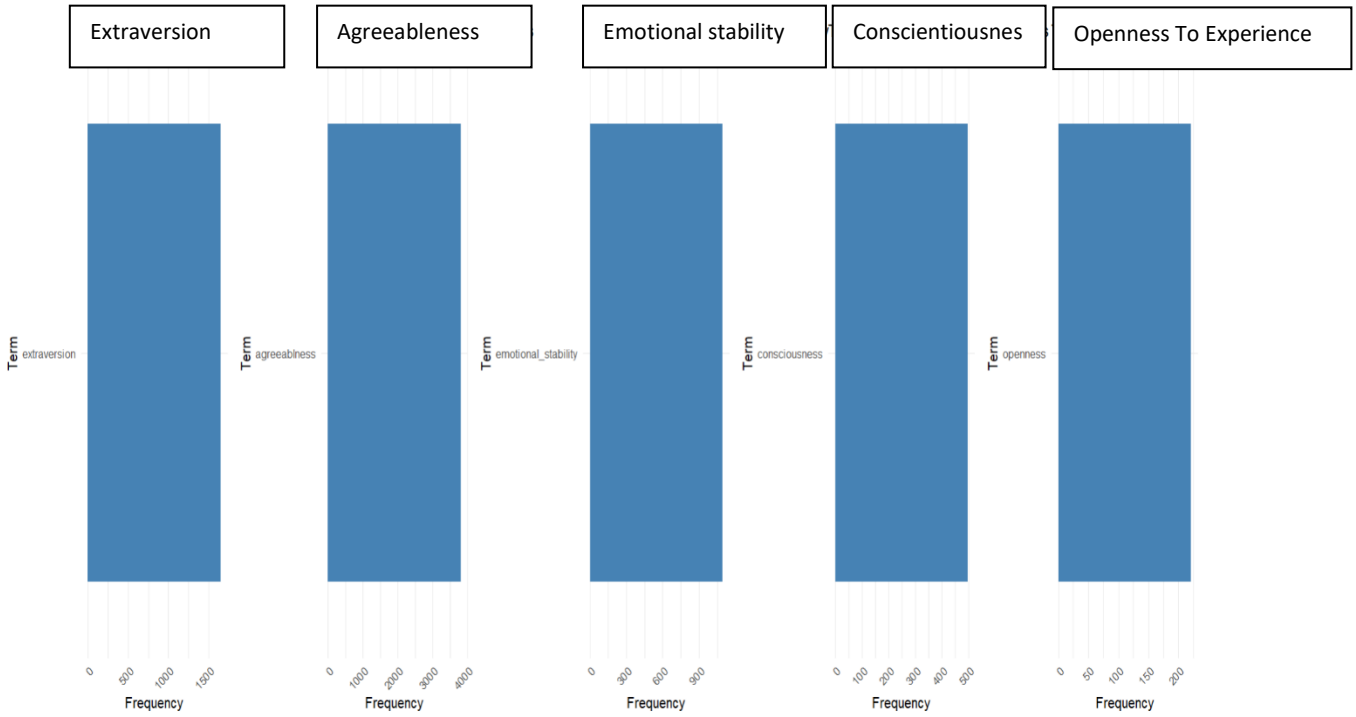


Figure 13: Five-Factor Dictionary Frequency

The initial step was to determine the frequency of terms under each personality classification before any topic extraction. Figures 12 and 13 above present the top 50 frequency of terms under each personality classification. Figure 12 investigates the three personality models: SAPI, English and the CPAI. It can be noted that SAPI is shown to have the highest frequency compared to English and CPAI. The reason is placed on the manner of term categorisation to avoid term overlap. Whereby SAPI was the default model, and terms found in the SAPI and the other model were not put there, remaining with SAPI having the highest number of traits. Figure 13 categorised Goldberg's model into the five factors that form part of the model.

Agreeableness has the highest frequency of terms, followed by Extraversion, Emotional stability, Conscientiousness and Openness. The frequency output of these top 50 words indicates a higher usage of terms that align with the agreeable trait. Personality research done in 36 cultures, including South Africa, found that Black South Africans were likely to score higher in Agreeableness and lower in extraversion and openness-to-experience when compared to other African and Asian cultures (Allik & McCrae, 2004; Vogt & Laher, 2009).

Compared to White South Africans, they scored lower on openness-to-experience, while White South Africans scored higher on extraversion and Agreeableness. This is translated in the term frequency under these clusters, with agreeableness-describing terms used more than openness-describing terms. It further speaks of the expression of personality traits rather than the actual personality (Vogt & Laher, 2009). Specific terms such as "rude" and "generous" speak to the expression of an agreeable individual and less so on their agreeable behaviour.

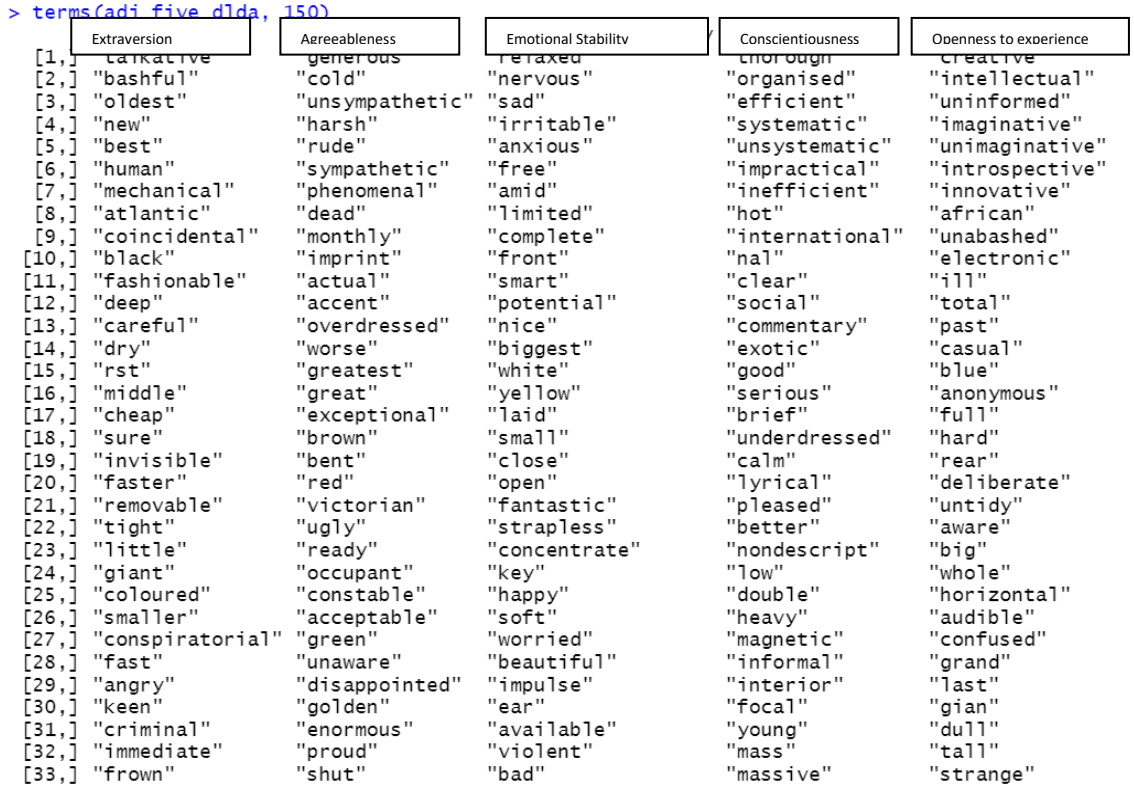
The understanding of the Openness to Experience personality cluster is still being explored. McCrae and Greenberg (2014) argue that other personality classifications can be easily described and expressed with certain personality traits, but with openness, attributing certain traits is often tricky. Even with the construction of the CPAI, there were initial doubts about the existence of openness among the Chinese people (F. Cheung et al., 2008). This research was founded on the Western understanding of openness. It has been seen that such an abstract expression of personality is more challenging to comprehend, making it difficult to replicate in the Sub-Saharan context (McCrae & Greenberg, 2014; Vogt & Laher, 2009). The smaller frequency under openness shows that minimal words are used in the representation and the expression of openness in the text.

5.4.2 Supervised Text Mining – LDA

A supervised LDA relies on the dictionaries it inserts into the model to guide the thematic extraction of the topics. In this case, each dictionary was either a model with its clusters (Goldberg) or different models in a single dictionary (SAPI, English and CPAI). Each cluster and model extracted words that aligned tended to fit with the topic based on occurrence-likelihood. The majority of the terms from the dictionaries were found in the text except for a few presented in Table 5. The model primarily extracts unigrams (singular words); the terms that could not be found were bi-grams (double words). It is mentioned that for the extraction of bi-grams, the co-occurrence needs to consider the frequency of words occurring together a couple of times before considering them as a single term (Zoya et al., 2021).

The terms that formed part of the models and clusters guided the selection of terms that frequently occurred under each pre-defined topic. The initial step was to evaluate the corpus containing all the words from the text. Though the terms were relatively located in the text, the noisy data was extracted, which made it difficult to gauge any underlying meaning behind the text, even with the help of the pre-defined personality terms. Zoya et al. (2021), when examining LDA and NMF Topic models for Urdu using Automatic Labelling, mention that though the LDA has capabilities such as document-topic distribution, it tends to produce

irrelevant topics. This is something that was expressed under the unsupervised topic modelling.



```
> terms(adj_five_dlda, 150)
      Extraversion    Agreeableness    Emotional Stability  Conscientiousness  Openness to experience
 [1,] "talkative"     "generous"       "relaxed"            "thorough"         "creative"
 [2,] "bashful"       "cold"           "nervous"            "organised"        "intellectual"
 [3,] "oldest"        "unsympathetic"  "sad"                "efficient"        "uninformed"
 [4,] "new"           "harsh"          "irritable"          "systematic"       "imaginative"
 [5,] "best"          "rude"           "anxious"            "unsystematic"     "unimaginative"
 [6,] "human"         "sympathetic"    "free"               "impractical"      "introspective"
 [7,] "mechanical"    "phenomenal"     "amid"               "inefficient"      "innovative"
 [8,] "atlantic"      "dead"           "limited"            "hot"              "african"
 [9,] "coincidental"  "monthly"        "complete"           "international"    "unabashed"
[10,] "black"         "imprint"        "front"              "nal"              "electronic"
[11,] "fashionable"   "actual"         "smart"              "clear"            "ill"
[12,] "deep"          "accent"         "potential"          "social"           "total"
[13,] "careful"       "overdressed"    "nice"               "commentary"       "past"
[14,] "dry"           "worse"          "biggest"            "exotic"           "casual"
[15,] "rst"           "greatest"       "white"              "good"             "blue"
[16,] "middle"        "great"          "yellow"             "serious"          "anonymous"
[17,] "cheap"         "exceptional"    "laid"               "brief"            "full"
[18,] "sure"          "brown"          "small"              "underdressed"     "hard"
[19,] "invisible"     "bent"           "close"              "calm"             "rear"
[20,] "faster"        "red"            "open"               "lyrical"          "deliberate"
[21,] "removable"     "victorian"      "fantastic"          "pleased"          "untidy"
[22,] "tight"         "ugly"           "strapless"          "better"           "aware"
[23,] "little"        "ready"          "concentrate"        "nondescript"      "big"
[24,] "giant"         "occupant"       "key"                "low"              "whole"
[25,] "coloured"      "constable"      "happy"              "double"           "horizontal"
[26,] "smaller"       "acceptable"     "soft"               "heavy"            "audible"
[27,] "conspiratorial" "green"         "worried"            "magnetic"         "confused"
[28,] "fast"          "unaware"        "beautiful"          "informal"         "grand"
[29,] "angry"         "disappointed"   "impulse"            "interior"         "last"
[30,] "keen"          "golden"         "ear"                "focal"            "gian"
[31,] "criminal"      "enormous"       "available"          "young"            "dull"
[32,] "immediate"     "proud"          "violent"            "mass"             "tall"
[33,] "frown"         "shut"           "bad"                "massive"          "strange"
```

Figure 14: Supervised LDA on Adjectives-only

When the supervised LDA was applied to the adjectives-only corpora, it yielded various probabilistically occurring terms for each pre-defined topic. Figure 14 showcases one of the topic selections of 33 words. This is not a seeded evaluation, so it forms one of the numerous probabilistic extractions per the rule of multi-nominal topic-word distribution (Zoya et al., 2021). It can be noted that Figure 13 included some pre-defined terms such as *"talkative"* and *"bashful"* in Extraversion, *"cold"* and *"generous"* from Agreeableness, *"relaxed"* and *"nervous"* from Emotional stability, *"thorough"* and *"organised"* from conscientiousness and *"creative"* and *"intellectual"* from Openness-to-Experience.

With the investigation of each topic, other terms were extracted that formed part of understanding personality clusters. These terms are subject to review as their thematic interpretation is based on the researchers' understanding of the personality clusters. Though topic modelling uses machine learning based on algorithmic processing of text, the assignment of relevancy of terms and themes still lies with the researcher (Hemmatian et al., 2019). As such, the classification of the terms is based on the basic understanding of the personality clusters and the trait theory research.

The current understanding of each cluster is that it has dimensions (low and high), and depending on temperament, one either falls on the high-end or low-end (Costa & McCrae, 1999; Lim, 2023; McCrae et al., 2000). In the Extraversion cluster, words *like "Fashionable", "Invisible",* and *"Frown"*. With the Agreeableness cluster, *"Phenomenal", "Greatest", "Acceptable"* and *"Proud"*. For Emotional stability, *"Worried", "Impulse", "Violent", "Bad",* and *"Free"*. On Conscientiousness, *"Clear", "Underdressed", "Serious", "Calm", "Pleased", "Informal"* and *"Nondescript"*. And lastly, with Openness-to-Experience, *"Unabashed", "Casual", "Deliberate", "Untidy",* and *"Strange"*.

These selected terms from Figure 14 capture a basic understanding of what each personality cluster means. They further manage to encompass the dimensions of each cluster. For example, the terms under extraversion can be interpreted as *"Fashionable",* being high in extraversion and *"Invisible",* and *"Frown",* being low in extraversion. Conscientiousness, *"Clear", "Serious", "Calm",* and *"Pleased"* would have a high score, while *"Underdressed", "Informal",* and *"Nondescript"* would be low scoring. Openness-to-Experience, *"Unabashed", "Untidy",* and *"Strange"* are high scores, while *"Casual"* and *"Deliberate"* are low scores.

The extraction of such terms using the supervised LDA shows that frameworks such as the Five Factor Model have a place in understanding personality, and this is assisted by techniques that do not rely on context to understand personality but rather traits and expression. When looking at other personality traits that the model extracted in comparison to openness, it can be seen that the traits are readily observable. Openness presents abstract representations of people's behaviour, such as *"Unabashed"* and *"Deliberate"*. The model adds to the description of openness but provides minimal physical attributes to the personality cluster. It has been noted by research that such a trait tends to be internal and experiential (McCrae & Greenberg, 2014).

When the cooccurrence of personality traits against pronouns was done, it was noted that a number of the personality traits from Goldberg's list under openness could not be found. Traits such as "*Introspective", "Uncreative", "Unimaginative", "Uninformed"* and *"Unintellectual"* were not present for males and females. It draws back to the cultural expression of such terms and the comprehension of openness as a cluster that tends to express the self. The openness-to-experience cluster tends to be in service of the self and the development of self-identity and individual interests which carries a more individualistic

view compared to agreeableness (Tesch & Cameron, 1987). The presentation of a lower frequency of openness to experience terms and higher agreeableness terms showcases the probabilistic nature of the LDA model in its ability to extract terms that have a high probability to cooccur in each topic.

As Hemmatian et al. (2019) warn, the topic modelling technique has potential but needs to be used within understanding of a topic. Such as inputting traits aligning with Openness to Experience for thematically aligning traits to be extracted. The supervised topic modelling technique showed that this thematic approach could be considered when working with large data and following the correct parameters. These parameters will currently be statistically based, and as such, the methodology still needs to be further explored and refined, following the type of data and the context in which it is being applied. The current exploration of topic modelling through LDA shows that it is a method that has flaws that researchers are often criticising (Hemmatian et al., 2019; Laureate et al., 2023; Okon et al., 2020), but it has shown the most flexibility and adaptability in various fields. It further confirmed some of the existing research outcomes on contextual understanding of the FFM.

5.5 Limitations

5.5.1 Large Corpus Issue

The text mining approach to personality investigation has shown great potential, especially in looking at the description of people and the gender cooccurrence exploration between pronouns and Goldberg's personality descriptors. However, the study came short of properly exploring personality distribution and clustering due to the large corpus of data. The study used 60 South African literary texts, each containing a minimum of 200 pages. This resulted in a large corpus of words, which made data analysis difficult, and specific model analyses took up to 3 days (some more depending on the complexity of the functions) to classify data.

5.5.2 Noisy Data

The use of the LDA method has been recognised to have limitations, such as the presence of noisy data and sparsity (Egger & Yu, 2022; Péladeau & Davoodi, 2018). This became a significant limitation in this study because the literary text contained words that belonged to some South African languages, such as Afrikaans, IsiZulu and IsiXhosa. South African authors are often bilingual and use some words to contextualise their writing. Some words that ended up in the topic extraction cannot be

found under stop words, which made it difficult to remove them individually as one would not know which word is used for what as part of text cleaning.

5.5.3 Optimal number of topics and interpretation

The selection of topics when using the LDA is often left up to the researchers' discretion given the knowledge of their field and what they are looking for in the text. There is no agreed-upon standard in the selection as the knowledge of the field is usually known by the researcher (Egger & Yu, 2022). This not only means the selection is limited to the researcher, but the interpretation of the topics as well lies with. This has a high chance of producing results that are biased in nature and need further validation. Furthermore, the topic selection can be limited in the amount of interpretation that can be yielded.

5.5.4 Exploration of Other Topic Models

While the LDA topic modelling approach is the most popular in classifying text and clustering thematic occurrences in each document, other models could have been explored. Models such as the Non-negative Matrix Factorization (NMF) are a non-probabilistic algorithmic approach to topic modelling that uses matrix factorisation under linear algebra (Egger & Yu, 2022). It allows for the clustering of high-dimensional data, which helps identify incoherent data.

5.6 Recommendations

5.6.1 Large Data Issue – Split Corpus

In utilising the approach of text mining for personality research, the corpus of words needs to be split. This is to ensure that the analyses performed run more smoothly and that the models do not face the issue of running for days at a time. Furthermore, an exploration into parallel LDA can be conducted. This is because the Parallel LDA is designed to counter the limitations of the LDA, which is a longer running time (Tian et al., 2019). As such, using this model for extensive data can be helpful for better analyses. Alternatively, a software change can also be fruitful as software such as Python has a better capacity for handling large corpora of data and has analytical packages that could prove helpful in extracting the relevant personality attributes and expanding better to contextual co-occurrence.

5.6.2 Noisy Data – Noise Controlling Approach

Literary text, like the one that forms the basis of the study, often contains stop words and words that clutter the analysis (noisy words). The stop words are often removed at the initial analysis phase as part of data cleaning; however, noisy words often remain. A Guided Noise Topic Model (GTM) is recommended for future personality research exploration. This is because the GTM is a semi-supervised model that allows the researcher to set seed for topics that they are interested in finding in each text, in the case of personality description (Churchill et al., 2022). Large datasets such as literary books or social media data often contain salient topics that can be picked up over the ones being investigated. This results in results that are not useful; as such, the employment of such noise-controlling models can combat that limitation.

5.6.3 Exploration of Other Topic Models – Comparing different models

As mentioned in section 5.6.2, using different models could result in a more robust topic selection and a better understanding of personality descriptions in natural text. Other models that could not be explored in the study include the Non-negative Factorization Model (NMF), which is a non-probabilistic approach to the topic model that manages to create matrices that cluster them using an algebraic approach (Egger & Yu, 2022). The use of such a model could give a different perspective on selecting a topic that is not as random as the LDA.

5.6.4 Future Research – Personality Construction, NLP, and Generative AI

The sphere of Natural Language Processing (NLP) is expanding quickly allowing for more algorithmic analysis of naturally occurring text. As it currently stands, researchers are using social media data as a means of understanding behaviour and personality construction with the limitation of dealing with large data. Generative AI opens the ability to build NLP models that are specialised in understanding personality construction and their context. The use of text mining has shown a layer of analysis when it comes to cooccurrences in understanding personality for different gender identities while topic modelling showed the thematic clustering to personality description. The further pursuit of the use of Generative AI has the potential to allow personality researchers to better norm the personality research while expanding the

understanding of personality constructs such as Openness-to-Experience in different cultures.

5.7 Conclusion

The discussion chapter integrates the methodological application of text mining techniques in understanding personality. The start of the chapter lays out the steps that were followed and the approaches of text mining in personality research. This was done by bringing about the foundation of data analysis, data cleaning, and setting up the text for more in-depth analyses, such as the pronoun-personality cooccurrence and unsupervised and supervised topic modelling. The chapter further examines the realities encountered in exploring the methodological approach and offers recommendations and future insights about text mining in personality research.

Psychology has relied on traditional (Qualitative/Quantitative) methodologies to expand personality research. Factor analysis is a way in which statistical and robust approaches are introduced to ensure the appropriate reproducibility of the data. Therefore, the challenge of handling extensive data arises with the increased social media engagement and use of virtual tools and mental health apps. Text mining, as a method founded in data science implementation of machine learning on natural language, offers a way in which personality learning can leverage these techniques to better understand personality construction. The continuously increasing number of techniques, software and algorithmic models indicates that personality research has a place in developing models and techniques that can help understand personality and the influences of context in its construction.

This paper investigated text mining to determine its suitability in personality research by exploring personality expression in South African literary texts. This meant a deep dive into the techniques of text mining from the collection of the data, type of data and cleaning right to the mining of the text, through the separation of wanted and unwanted text and modelling the text. These techniques showed the usability of text mining at different stages and the exploration of personality expression in those stages. For example, the POS tagging managed to show a way in which personality expression can be understood alongside presented gender identities. The term

frequency showed the most used terms in the text, which allowed for the exploration of the frequency of terms in each personality cluster. This gave insight into the type of frequent personality expression common in the text.

Topic modelling is used to discover themes that ought to make the text meaningful. Such an approach showed its usefulness when used on text that shares a similar theme. Though it came with limitations when used on large text of various thematic undertones, it was insightful in determining the thematic alignment of terms when a supervised topic model was used. It assisted in determining the kinds of personality expressions associated with each personality structure. The supervised LDA model showed the potential when used for further investigation of personality construction. This initial discovery can help build up the understanding of personality expression in the South African context and help form a better comprehension of personality cluster formulation. Natural Language Processing (NLP) through text mining brings a new light when analysing and quantifying qualitative data. As such, future research on using NLP, creating methodological guidelines, creating contextual norms, and integrating newer methods (Generative AI and LLMs) in personality research is needed.

References

Abrahams, F., & Mauer, K. F. (1999). Qualitative and Statistical Impacts of Home Language on Responses to the Items of the Sixteen Personality Factor Questionnaire (16PF) in South Africa. *South African Journal of Psychology*, *29*(2), 76–86. https://doi.org/10.1177/008124639902900204

Akbari, M., Seydavi, M., Spada, M. M., Mohammadkhani, S., Jamshidi, S., Jamaloo, A., & Ayatmehr, F. (2021). The Big Five personality traits and online gaming: A systematic review and meta-analysis. *Journal of Behavioral Addictions*, *10*(3), 611–625. https://doi.org/10.1556/2006.2021.00050

Allik, J., & McCrae, R. R. (2004). Toward a Geography of Personality Traits: Patterns of Profiles across 36 Cultures. *Journal of Cross-Cultural Psychology*, *35*(1), 13–28. https://doi.org/10.1177/0022022103260382

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), i–171. https://doi.org/10.1037/h0093360

Andersson, I., Persson, J., & Kajonius, P. (2022). Even the stars think that I am superior: Personality, intelligence and belief in astrology. *Personality and Individual Differences*, *187*, 111389. https://doi.org/10.1016/j.paid.2021.111389

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Management*, *50*(3), 329–351. https://doi.org/10.1111/radm.12408

Archer, R. P., & Smith, S. R. (2011). *Personality Assessment*. Routledge.

Babu, D. M. S., Ali, M. A. A., & Rao, M. A. S. (2018). A Study on Information Retrieval Methods in Text Mining. *International Journal of Engineering Research & Technology*, *2*(15). https://doi.org/10.17577/IJERTCONV2IS15028

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, J. J. A., Fleeson, W., Grafton, B., Jayawickreme, E., Kurzius, E., MacLeod,

C., Miller, L. C., Read, S. J., Roberts, B., Robinson, M. D., Wood, D., & Wrzus, C. (2017). Integrating Personality Structure, Personality Process, and Personality Development. *European Journal of Personality*, *31*(5), 503–528. https://doi.org/10.1002/per.2115

Beck, E. D., & Jackson, J. J. (2021). Chapter 4—Within-person variability. In J. F. Rauthmann (Ed.), *The Handbook of Personality Dynamics and Processes* (pp. 75–100). Academic Press. https://doi.org/10.1016/B978-0-12-813995-0.00004-2

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), Article 30. https://doi.org/10.21105/joss.00774

Bhawuk, D. P. (2017). Individualism and collectivism. *The International Encyclopedia of Intercultural Communication*, *2*, 920–929.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*.

Bonch-Osmolovskaya, A., & Skorinkin, D. (2017). Text mining War and Peace: Automatic extraction of character traits from literary pieces. *Digital Scholarship in the Humanities*, *32*(suppl_1), i17–i24. https://doi.org/10.1093/llc/fqw052

Bonner, C., Tuckerman, J., Kaufman, J., Costa, D., Durrheim, D. N., Trevena, L., Thomas, S., & Danchin, M. (2021). Comparing inductive and deductive analysis techniques to understand health service implementation problems: A case study of childhood vaccination barriers. *Implementation Science Communications*, *2*(1), 100. https://doi.org/10.1186/s43058-021-00202-0

Boumans, J. W., & Trilling, D. (2016). Taking Stock of the Toolkit. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Bowen, G. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*, 27–40. https://doi.org/10.3316/QRJ0902027

Boyle, G. J., Stankov, L., Martin, N. G., Petrides, K. V., Eysenck, M. W., & Ortet, G. (2016). Hans J. Eysenck and Raymond B. Cattell on intelligence and personality. *Personality and Individual Differences*, *103*, 40–47. https://doi.org/10.1016/j.paid.2016.04.029

Buecker, S., Maes, M., Denissen, J. J. A., & Luhmann, M. (2020). Loneliness and the Big Five Personality Traits: A Meta-analysis. *European Journal of Personality*, *34*(1), 8–28. https://doi.org/10.1002/per.2229

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, *38*(4), 476–506. https://doi.org/10.1037/h0054116

Cattell, R. B. (1965). *The scientific analysis of personality* (p. 399). Penguin Books.

Cattell, R. B., & Cattell, H. E. (1995). Personality Structure and the New Fifth Edition of the 16PF. *Educational and Psychological Measurement*, *55*(6), 926–937. https://doi.org/10.1177/0013164495055006002

Cattell, R. B., & Krug, S. E. (1986). The Number of Factors in the 16PF: A Review of the Evidence with Special Emphasis on Methodological Problems. *Educational and Psychological Measurement*, *46*(3), 509–522. https://doi.org/10.1177/0013164486463002

Celardo, L., & Everett, M. G. (2020). Network text analysis: A two-way classification approach. *International Journal of Information Management*, *51*, 102009. https://doi.org/10.1016/j.ijinfomgt.2019.09.005

Cheng, C.-H., & Chen, H.-H. (2019). Sentimental text mining based on an additional features method for text classification. *PLOS ONE*, *14*(6), e0217591. https://doi.org/10.1371/journal.pone.0217591

Cheung, C., Lee, M. K. O., & Rabjohn, N. (2008). The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities. *Internet Research*, *18*(3), 229–247. https://doi.org/10.1108/10662240810883290

Cheung, F., Cheung, S. F., & Fan, W. (2013). From Chinese to cross-cultural inventory: A combined emic-etic approach to the study of personality in culture. In *Advances in Culture and Psychology: Volume 3*. OUP USA.

Cheung, F., Cheung, S. F., Zhang, J., Leung, K., Leong, F., & Yeh, K.-H. (2008). Relevance of Openness as a Personality Dimension in Chinese Culture: Aspects of its Cultural Relevance. *Journal of Cross-Cultural Psychology*, *39*, 81–108. https://doi.org/10.1177/0022022107311968

Cheung, F., Cheung, S. F., Wada, S., & Zhang, J. (2003). Indigenous Measures of Personality Assessment in Asian Countries: A Review. *Psychological Assessment*, *15*(3), 280–289. https://doi.org/10.1037/1040-3590.15.3.280

Cheung, F., & Leung, K. (1998). Indigenous Personality Measures: Chinese Examples. *Journal of Cross-Cultural Psychology*, *29*(1), 233–248. https://doi.org/10.1177/0022022198291012

Cheung, F., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist*, *66*(7), 593. https://doi.org/10.1037/a0022389

Cheung, F., Leung, K., Fan, R. M., Song, W.-Z., Zhang, J.-X., & Zhang, J.-P. (1996). Development of the Chinese Personality Assessment Inventory. *Journal of Cross-Cultural Psychology*, *27*(2), 181–199. https://doi.org/10.1177/0022022196272003

Cheung, P. C., Conger, A. J., Hau, K.-T., Lew, W. J. F., & Lau, S. (1992). Development of the Multi-Trait Personality Inventory (MTPI): Comparison Among Four Chinese

Populations. *Journal of Personality Assessment*, *59*(3), 528–551. https://doi.org/10.1207/s15327752jpa5903_8

Chico, E., & Lorenzo-Seva, U. (2006). Belief in Astrology Inventory: Development and Validation. *Psychological Reports*, *99*(3), 851–863. https://doi.org/10.2466/PR0.99.3.851-863

Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, *8*(1), 68. https://doi.org/10.1186/s40537-021-00459-1

Churchill, R., Singh, L., Ryan, R., & Davis-Kean, P. (2022). A Guided Topic-Noise Model for Short Texts. *Proceedings of the ACM Web Conference 2022*, 2870–2878. https://doi.org/10.1145/3485447.3512007

Ciccarelli, S. K. (2006). *Psychology*. Upper Saddle River, NJ Pearson / Prentice Hall. http://archive.org/details/psychology0000cicc_s6n4

Cooper, S. (2014). South African psychology 20 years into democracy. *South African Journal of Psychology*, *44*(3), 261–266. https://doi.org/10.1177/0081246314537176

Costa, P., & McCrae, R. (2012). The Five-Factor Model, Five-Factor Theory, and Interpersonal Psychology. *Handbook of Interpersonal Psychology: Theory, Research, Assessment, and Therapeutic Interventions*, 91–104. https://doi.org/10.1002/9781118001868.ch6

Costa, P., & McCrae, R. R. (1999). A five-factor theory of personality. *The Five-Factor Model of Personality: Theoretical Perspectives*, *2*, 51–87.

Costa, P., Terracciano, A., & McCrae, R. (2001). Gender Differences in Personality Traits Across Cultures: Robust and Surprising Findings. *Journal of Personality and Social Psychology, 81*, 322–331. https://doi.org/10.1037//0022-3514.81.2.322

CPAI Background. (n.d.). Retrieved 4 August 2023, from https://cpaiweb.psy.cuhk.edu.hk/English/AboutCPAI.html

De Fruyt, F., & Wille, B. (2013). *Cross-cultural issues in personality assessment*.

De Raad, B., & Mlacic, B. (2015). *The Lexical Foundation of the Big Five-Factor Model* (p. 31). https://doi.org/10.1093/oxfordhb/9780199352487.013.12

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), Article 11. https://doi.org/10.1038/s44159-023-00241-5

Dexter, S. (2017, January 24). *Text Analytics: A Primer | GreenBook*. https://www.greenbook.org/mr/market-research-leaders/text-analytics-a-primer/

Diener, E., Lucas, R. E., & Cummings, J. A. (2019). Personality Traits. In *Introduction to Psychology*. University of Saskatchewan Open Press. https://openpress.usask.ca/introductiontopsychology/chapter/personality-traits/

Doremus, C. (2020). Trait Theory of Allport. In *The Wiley Encyclopedia of Personality and Individual Differences, Models and Theories*. John Wiley & Sons.

Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, *7*, 886498. https://doi.org/10.3389/fsoc.2022.886498

Eysenck, H. J. (1953). *The structure of human personality* (pp. xix, 348). Methuen.

Eysenck, H. J. (1991). Dimensions of personality: The biosocial approach to personality. In *Explorations in temperament: International perspectives on theory and measurement* (pp. 87–103). Plenum Press. https://doi.org/10.1007/978-1-4899-0643-4_7

Gaikwad, S., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications*, *85*(17), 42–45. https://doi.org/10.5120/14937-3507

Gan, J., & Qi, Y. (2021). Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. *Entropy*, *23*(10). https://doi.org/10.3390/e23101301

Gelfand, M., Bhawuk, D., Nishii, L., & Bechtold, D. (2004). *Individualism and Collectivism* (pp. 437–512). https://doi.org/10.1002/9781118783665.ieicc0107

Gillis, J., & Boyle, G. (2018). Factor Analysis of Trait name. Revisiting Cattell (1943). In *Personality and Individual Differences: Revisiting the Classic Studies*. SAGE.

Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '11*, 253. https://doi.org/10.1145/1979742.1979614

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Government Gazette. (1978). *Copyright Act 98 of 1978 | South African Government*. https://www.gov.za/documents/copyright-act-16-apr-2015-0942

Government Gazette. (1998). *REPUBLIC OF SOUTH AFRICA. Employment Equity Act, 1998*. https://www.ilo.org/dyn/natlex/docs/WEBTEXT/51169/65139/E98ZAF01.htm

Government Gazette. (2013). *Protection of Personal Information Act*.

Grzybowski, S., Wyczesany, M., Cichecka, H., & Tokarska, A. (2021). The Words of Affectivity. Affect, Category, and Social Evaluation Norms for 400 Polish Adjectives. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.683012

Gunter, B. (2019). *Personality Traits in Online Communication*. Routledge.

Gurven, M., Massenkoff, M., & Kaplan, H. (2013). How Universal Is the Big Five? Testing the Five-Factor Model of Personality Variation Among Forager–Farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology*, *104*(2), 354–370. https://doi.org/DOI: 10.1037/a0030841

Hemmatian, B., Sloman, S. J., Cohen Priva, U., & Sloman, S. A. (2019). Think of the consequences: A decade of discourse about same-sex marriage. *Behavior Research Methods*, *51*(4), 1565–1585. https://doi.org/10.3758/s13428-019-01215-3

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

Hill, C., Hlahleni, M., & Legodi, L. (2021). Validating Indigenous Versions of the South African Personality Inventory. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.556565

Hill, C., Nel, J. A., Vijver, F. J. R. van de, Meiring, D., Valchev, V. H., Adams, B. G., & Bruin, G. P. de. (2013). Developing and testing items for the South African Personality Inventory (SAPI). *SA Journal of Industrial Psychology*, *39*(1), Article 1. https://doi.org/10.4102/sajip.v39i1.1122

Hiran, K. K., Jain, R. K., Lakhwani, D. K., & Doshi, D. R. (2021). *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*. BPB Publications.

Holtgraves, T. M. (2014). *The Oxford Handbook of Language and Social Psychology*. Oxford University Press.

Ignatow, G., & Mihalcea, R. (2017). *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. SAGE Publications.

Jeremy, N. H., Prasetyo, C., & Suhartono, and D. (2019). Identifying Personality Traits for Indonesian User from Twitter Dataset. *International Journal of Fuzzy Logic and Intelligent Systems*, *19*(4), 283–289. https://doi.org/10.5391/IJFIS.2019.19.4.283

John, O. P., & Robins, R. W. (2021). *Handbook of Personality, Fourth Edition*. Guilford Publications.

Jung, C. (2016). *Psychological Types*. Routledge.

Jung, C. G., & Hull, R. F. C. (1976). *Collected Works of C.G. Jung, Volume 6: Psychological Types*. Princeton University Press.

Kahya-Özyirmidokuz, E. (2014). *Mining Unstructured Turkish Economy News Articles*. *16*. https://doi.org/10.1016/S2212-5671(14)00809-0

Kamal, R., & Saxena, P. (2019). *Big Data Analytics... Web Social Network Analytics*. Scribd. https://www.scribd.com/document/601773408/BDACh02L01Hadoop

Katigbak, M. S., Church, A. T., Guanzon-Lapeña, Ma. A., Carlota, A. J., & del Pilar, G. H. (2002). Are indigenous personality dimensions culture specific? Philippine inventories and the five-factor model. *Journal of Personality and Social Psychology*, *82*(1), 89–101. https://doi.org/10.1037/0022-3514.82.1.89

Kaur, K. M., Malé, P.-J. G., Spence, E., Gomez, C., & Frederickson, M. E. (2019). Using text-mined trait data to test for cooperate-and-radiate co-evolution between ants and plants. *PLOS Computational Biology*, *15*(10), e1007323. https://doi.org/10.1371/journal.pcbi.1007323

Kavirayani, K. (2018). Historical perspectives on personality – The past and current concept: The search is not yet over. *Archives of Medicine and Health Sciences*, *6*(1), 180. https://doi.org/10.4103/amhs.amhs_63_18

Kwartler, T. (2017). *Text Mining in Practice with R*. Wiley.

Laher, S. (2015). Exploring the utility of the CPAI-2 in a South African sample: Implications for the FFM. *Personality and Individual Differences*, *81*, 67–75. https://doi.org/10.1016/j.paid.2014.12.010

Laher, S., Cheung, F., & Zeinoun, P. (2020). *Gender and Personality Research in Psychology: The Need for Intersectionality* (pp. 167–178). https://doi.org/10.1017/9781108561716.016

Laher, S., & Cockcroft, K. (2014). Psychological assessment in post-apartheid South Africa: The way forward. *South African Journal of Psychology*, *44*(3), 303–314. https://doi.org/10.1177/0081246314533634

Laher, S., & Dockrat, S. (2019). The five-factor model and individualism and collectivism in South Africa: Implications for personality assessment. *African Journal of Psychological Assessment*, *1*(0), Article 0. https://ajopa.org/index.php/ajopa/article/view/4

Laureate, C. D. P., Buntine, W., & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, *56*(12), 14223–14255. https://doi.org/10.1007/s10462-023-10471-x

Lee, S. (2017). Rethinking the relationship between pronoun-drop and individualism with Bayesian multilevel models. *Journal of Language Evolution*, *2*(2), 188–200. https://doi.org/10.1093/jole/lzx003

Li, X., Wang, Y., Zhang, A., Li, C., Chi, J., & Ouyang, J. (2018). Filtering out the noise in short text topic modeling. *Information Sciences*, *456*, 83–96. https://doi.org/10.1016/j.ins.2018.04.071

Lim, A. (2023, December 20). *Big 5 Personality Traits: The 5-Factor Model of Personality*. https://www.simplypsychology.org/big-five-personality.html

Lin, C., He, Y., Pedrinaci, C., & Domingue, J. (2012). Feature LDA: A Supervised Topic Model for Automatic Detection of Web API Documentations from the Web. In P. Cudré-Mauroux, J. Heflin, E. Sirin, T. Tudorache, J. Euzenat, M. Hauswirth, J. X. Parreira, J. Hendler, G. Schreiber, A. Bernstein, & E. Blomqvist (Eds.), *The Semantic Web – ISWC 2012* (pp. 328–343). Springer. https://doi.org/10.1007/978-3-642-35176-1_21

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, *5*(1), 1608. https://doi.org/10.1186/s40064-016-3252-8

Liu, Q., & Wu, Y. (2012). *Supervised Learning*.

Livaniene, V., & De Raad, B. (2017). The factor structure of Lithuanian personality-descriptive adjectives of the highest frequency of use. *International Journal of Psychology*, *52*(6), 453–462. https://doi.org/10.1002/ijop.12247

Malkappagol, D. R. G. (2018). *EFFECT OF EMOTIONAL MATURITY AND PERSONALITY ON WELL-BEING AMONG TEACHERS*. Laxmi Book Publishing.

Matthews, G., Deary, I. J., & Whiteman, M. C. (2009). *Personality Traits*. Cambridge University Press.

Mcauliffe, J., & Blei, D. (2007). Supervised Topic Models. *Advances in Neural Information Processing Systems*, *20*.

https://proceedings.neurips.cc/paper_files/paper/2007/hash/d56b9fc4b0f1be8871f5e1c40c0067e7-Abstract.html

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509–516. https://doi.org/10.1037/0003-066X.52.5.509

McCrae, R. R., Costa, P. T., Ostendorf, F., Angleitner, A., Hrebícková, M., Avia, M. D., Sanz, J., Sánchez-Bernardos, M. L., Kusdil, M. E., Woodfield, R., Saunders, P. R., & Smith, P. B. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology*, *78*(1), 173–186. https://doi.org/10.1037//0022-3514.78.1.173

McCrae, R. R., & Greenberg, D. M. (2014). Openness to Experience. In D. K. Simonton (Ed.), *The Wiley Handbook of Genius* (1st ed., pp. 222–243). Wiley. https://doi.org/10.1002/9781118367377.ch12

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McCrae, & Terracciano. (2005). Universal Features of Personality Traits From the Observer's Perspective: Data From 50 Cultures. *Journal of Personality and Social Psychology*, *88*, 547–561. https://doi.org/10.1037/0022-3514.88.3.547

Meyer, L. H. and G.]. (1998). The Importance of Teaching and Learning Personality Assessment. In *Teaching and Learning Personality Assessment*. Routledge.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, G. B., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*. https://doi.org/10.1126/science.1199644

Mollaret, P. (2009). Using Common Psychological Terms to Describe Other People: From Lexical Hypothesis to Polysemous Conception. *Theory & Psychology*, *19*(3), 315–334. https://doi.org/10.1177/0959354309104157

Monroe, B. (2021). *An Introduction to Text as Data with quanteda*. https://burtmonroe.github.io/TextAsDataCourse/Tutorials/TADA-IntroToQuanteda.html

Morton, N., Hill, C., Meiring, D., & de Beer, L. T. (2019). Investigating the factor structure of the South African Personality Inventory—English version. *SA Journal of Industrial Psychology*, *45*(1), 1–13. https://doi.org/10.4102/sajip.v45i0.1556

Murshed, B. A. H., Mallappa, S., Abawajy, J., Saif, M. A. N., Al-ariki, H. D. E., & Abdulwahab, H. M. (2023). Short text topic modelling approaches in the context of big data: Taxonomy, survey, and analysis. *Artificial Intelligence Review*, *56*(6), 5133–5260. https://doi.org/10.1007/s10462-022-10254-w

Nel, J. A., Valchev, V. H., Rothmann, S., Vijver, F. J. R., Meiring, D., & Bruin, G. P. (2012). Exploring the Personality Structure in the 11 Languages of South Africa. *Journal of Personality*, *80*(4), 915–948. https://www.academia.edu/11882015/Exploring_the_Personality_Structure_in_the_11_Languages_of_South_Africa

Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., & Winters, J. (2020). How We Do Things With Words: Analyzing Text as Social and Cultural Data. *Frontiers in Artificial Intelligence*, *3*. https://www.frontiersin.org/article/10.3389/frai.2020.00062

O'Connor, B., Bamman, D., & Smith, N. A. (2011). Computational Text Analysis for Social Science: Model Assumptions and Complexity. *Public Health*, *41*(43), 7.

Okon, E., Rachakonda, V., Hong, H. J., Callison-Burch, C., & Lipoff, J. B. (2020). Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. *Journal of the American Academy of Dermatology*, *83*(3), 803–808. https://doi.org/10.1016/j.jaad.2019.07.014

Osman, M. (2022). *Wild and Interesting Facebook Statistics and Facts (2023)*. Kinsta®. https://kinsta.com/blog/facebook-statistics/

Paranyushkin, D. (2019). InfraNodus: Generating Insight Using Text Network Analysis. *The World Wide Web Conference*, 3584–3589. https://doi.org/10.1145/3308558.3314123

Péladeau, N., & Davoodi, E. (2018). *Comparison of Latent Dirichlet Modeling and Factor Analysis for Topic Extraction: A Lesson of History*. http://hdl.handle.net/10125/49965

Pennebaker, J., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. *University of Texas at Austin*, 26.

Pennebaker, J., & King, L. (2000). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*, 1296–1312. https://doi.org/10.1037//0022-3514.77.6.1296

Pincus, A. L. (2010). Five-Factor Model of Personality. In *The Corsini Encyclopedia of Psychology* (pp. 1–2). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470479216.corpsy0362

Posel, D., Hunter, M., & Rudwick, S. (2020). Revisiting the prevalence of English: Language use outside the home in South Africa. *Journal of Multilingual and Multicultural Development*, 1–13. https://doi.org/10.1080/01434632.2020.1778707

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). *Mastering Modern Psychological Testing: Theory and Methods*. Springer Nature.

Roivainen, E. (2013). Frequency of the use of English personality adjectives: Implications for personality theory. *Journal of Research in Personality*, *47*(4), 417–420. https://doi.org/10.1016/j.jrp.2013.04.004

Roivainen, E. (2022). Age of Acquisition of Personality Terms: Implications for Personality Theory. *Europe's Journal of Psychology*, *18*(2), 132–141. https://doi.org/10.5964/ejop.2987

Sathya, R., & Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38.

Saucier, G. (2008). Measures of the personality factors found recurrently in human lexicons. In *The SAGE handbook of personality theory and assessment, Vol 2: Personality measurement and testing* (pp. 29–54). Sage Publications, Inc. https://doi.org/10.4135/9781849200479.n2

Schultz, D., & Schultz, S. (2004). *Theories of Personality*. Cengage Learning.

Schultz, D., & Schultz, S. (2016). *Theories of Personality*. Cengage Learning.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc.

Singh, J. K., Misra, G., & De Raad, B. (2013). Personality Structure in the Trait Lexicon of Hindi, a Major Language Spoken in India. *European Journal of Personality*, *27*(6), 605–620. https://doi.org/10.1002/per.1940

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2018). Personality Predictions Based on User Behavior on the Facebook Social Media Platform. *IEEE Access*, *6*, 61959–61969. https://doi.org/10.1109/ACCESS.2018.2876502

Talburt, J. (1985). *The Flesch Index: An Easily Programmable Readability Analysis Algorithm*. 114–122. https://doi.org/10.1145/10563.10583

Talib, R., Kashif, M., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, *7*. https://doi.org/10.14569/IJACSA.2016.071153

Tavakol, M., & Wetzel, A. (2020). Factor Analysis: A means for theory and instrument development in support of construct validity. *International Journal of Medical Education*, *11*, 245–247. https://doi.org/10.5116/ijme.5f96.0f4a

Tesch, S. A., & Cameron, K. A. (1987). Openness to Experience and Development of Adult Identity. *Journal of Personality*, *55*(4), 615–630. https://doi.org/10.1111/j.1467-6494.1987.tb00455.x

Tett, R. P., Toich, M. J., & Ozkum, S. B. (2021). Trait Activation Theory: A Review of the Literature and Applications to Five Lines of Personality Dynamics Research. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*(1), 199–233. https://doi.org/10.1146/annurev-orgpsych-012420-062228

Thalmayer, A. G., Job, S., Shino, E. N., Robinson, S. L., & Saucier, G. (2021). ǂŪsigu: A mixed-method lexical study of character description in Khoekhoegowab. *Journal of Personality and Social Psychology*, *121*(6), 1258–1283. https://doi.org/10.1037/pspp0000372

Thalmayer, A. G., Saucier, G., Shino, E. N., & Job, S. (2021). The Khoekhoegowab Personality Inventory: The Comparative Validity of a Locally Derived Measure of

Traits. *Frontiers in Psychology*, *12*, 694205. https://doi.org/10.3389/fpsyg.2021.694205

Thalmayer, Saucier, G., & Rotzinger, J. S. (2022). Absolutism, Relativism, and Universalism in Personality Traits Across Cultures: The Case of the Big Five. *Journal of Cross-Cultural Psychology*, *53*(7–8), 935–956. https://doi.org/DOI: 10.1177/00220221221111813

Thielmann, A., Weisser, C., Krenz, A., & Säfken, B. (2020). *Unsupervised Document Classification integrating Web Scraping, One-Class SVM and LDA Topic Modelling*.

Thisarani, M. (2021, June 25). Personality Detection and Prediction Using Natural Language Processing. *Medium*. https://ygsl-crew.medium.com/personality-detection-and-prediction-using-natural-language-processing-c2cd5cb4a2c7

Tian, Z., Yokoyama, H., & Araki, T. (2019). Parallel Latent Dirichlet Allocation Using Vector Processors. *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 1548–1555. https://doi.org/10.1109/HPCC/SmartCity/DSS.2019.00213

Tohver, G. C. (2020). Eysenck Giant Three. In *The Wiley Encyclopedia of Personality and Individual Differences* (pp. 155–159). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118970843.ch203

Triandis, H. C. (2001). Individualism-Collectivism and Personality. *Journal of Personality*, *69*(6), 907–924. https://doi.org/10.1111/1467-6494.696169

Triandis, H. C. (2018). *Individualism And Collectivism*. Routledge.

Uher, J. (2013). Personality Psychology: Lexical Approaches, Assessment Methods, and Trait Concepts Reveal Only Half of the Story—Why it is Time for a Paradigm Shift.

Integrative Psychological and Behavioral Science, *47*(1), 1–55. https://doi.org/10.1007/s12124-013-9230-6

Valchev, V. (2012). *Personality and Culture in South Africa*. Ridderprint.

Valchev, V., van de Vijver, F. J. R., Alewyn Nel, J., Rothmann, S., Meiring, D., & de Bruin, G. P. (2011). Implicit Personality Conceptions of the Nguni Cultural-Linguistic Groups of South Africa. *Cross-Cultural Research*, *45*(3), 235–266. https://doi.org/10.1177/1069397111402462

Vogt, L., & Laher, S. (2009). The five factor model of personality and individualism/collectivism in South Africa: An exploratory study. *Psychology in Society*, *37*, 39–54. http://www.scielo.org.za/scielo.php?script=sci_abstract&pid=S1015-60462009000200003&lng=en&nrm=iso&tlng=en

Wang, W., Jiang, X., Tian, S., Liu, P., Dang, D., Su, Y., Lookman, T., & Xie, J. (2022). Automated pipeline for superalloy data by text mining. *Npj Computational Materials*, *8*(1), Article 1. https://doi.org/10.1038/s41524-021-00687-2

Weiner, I. B., & Greene, R. L. (2017). *Handbook of Personality Assessment*. John Wiley & Sons.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, *11*(4), 245–265. https://doi.org/10.1080/19312458.2017.1387238

Wijffels, J., Straka, M., Straková, J., & BNOSAC (2020). *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit.* R package version 0.8.4. https://CRAN.R-project.org/package=udpipe

Wood, D. (2015). Testing the lexical hypothesis: Are socially important traits more densely reflected in the English lexicon? *Journal of Personality and Social Psychology*, *108*(2), 317–335. https://doi.org/10.1037/a0038343

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

Zanini, N., & Dhawan, V. (2015). *Text Mining: An introduction to theory and some applications*. *19*.

Zeinoun, P., Daouk-Öyry, L., Choueiri, L., & van de Vijver, F. J. R. (2018). Arab-Levantine personality structure: A psycholexical study of modern standard Arabic in Lebanon, Syria, Jordan, and the West Bank. *Journal of Personality*, *86*(3), 397–421. https://doi.org/10.1111/jopy.12324

Zell, E., & Lesick, T. L. (2022). Big five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality*, *90*(4), 559–573. https://doi.org/10.1111/jopy.12683

Zoya, Latif, S., Shafait, F., & Latif, R. (2021). Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling. *IEEE Access*, *9*, 127531–127547. https://doi.org/10.1109/ACCESS.2021.3112620

Appendix – Corpus full details

Text	Characters	Sentences	Tokens	Types	Punctuations	Numbers	Symbols	URLs	Tags	Emojis
1	599543	11064	135511	8960	28235	328	10	33	1	0
2	229227	3120	50421	5358	8494	26	2	0	0	0
3	1328638	19275	282942	17843	37220	554	18	3	0	0
4	302968	5368	70568	6501	13293	39	0	0	0	0
5	529964	8285	112031	8665	18013	36	1	0	0	0
6	458289	5948	100639	7787	16680	30	1	0	1	0
7	467490	7679	98807	8234	17242	92	1	0	0	0
8	283600	4634	61132	6178	9794	129	2	2	0	0
9	580681	10821	131363	8611	27881	392	20	9	0	0
10	253386	3466	55412	6164	7982	47	4	1	0	0
11	259626	2903	55105	5916	7632	18	1	0	0	0
12	452876	5100	103142	5278	16694	73	0	0	0	0
13	324755	3644	68458	8380	9276	292	34	0	0	0
14	358262	6569	81837	7741	15523	43	1	0	0	0
15	273960	3271	57001	8086	7339	207	11	0	0	0
16	272494	2411	54524	7140	7351	27	1	0	0	0
17	620703	10333	137105	9920	26411	270	3	2	0	0
18	649521	9593	143852	9844	22235	185	70	1	1	0
19	601510	8468	130283	9644	19560	250	12	2	0	0
20	476151	5867	97219	8151	15163	208	1	0	0	0
21	260560	2131	55162	6779	8132	81	1	0	0	0
22	916848	14075	208529	13811	36457	474	3	24	1	0
23	352487	4403	76890	6813	10374	34	1	0	0	0
24	198192	2702	42007	4300	5795	4	16	0	0	0
25	528167	10331	116582	8803	20322	179	11	0	1	0
26	1223939	12483	256891	14938	36228	101	2	0	0	0
27	319279	8355	72661	6336	11660	34	1	0	0	0
28	667590	11530	141468	12212	27266	307	4	1	0	0
29	507979	7061	106580	8972	15919	81	1	0	0	0

30	153258	2318	33435	4080	4636	80	2	0	0	0
31	452070	4192	93382	9583	11915	22	1	0	0	0
32	513643	9654	109965	8594	17459	147	2	0	0	0
33	300708	4049	64203	6617	9222	41	8	0	0	0
34	968830	12166	208726	14283	29071	163	2	0	0	0
35	416803	6867	94404	8089	16598	28	1	0	0	0
36	246046	3380	53513	4756	7962	85	2	0	0	0
37	1875498	22696	391891	17091	54535	81	1	0	0	0
38	369560	6629	81969	6827	12769	52	6	0	16	0
39	436011	8932	103768	6672	20857	92	1	0	0	0
40	2718364	26308	561393	21654	85218	560	31	0	0	0
41	410446	5294	85535	10058	10151	287	16	1	1	0
42	370702	6911	85383	6002	16112	58	1	0	0	0
43	562914	8427	116987	8355	16594	44	2	1	0	0
44	652289	11239	141421	10343	26146	366	8	17	0	0
45	234165	2210	48133	6038	7542	344	1	1	0	0
46	467258	8311	108536	6840	22647	121	1	0	0	0
47	404997	7337	93110	6692	17461	76	1	0	0	0
48	497502	6228	111100	8492	22813	108	1	0	0	0
49	506488	10838	115827	8390	22823	135	19	0	0	0
50	424194	6050	89621	7933	11465	38	2	1	0	0
51	174460	2565	38557	4427	7529	91	3	2	0	0
52	271299	5896	63341	4828	11445	302	4	0	0	0
53	327168	5952	75854	5015	14544	21	1	0	0	0
54	377804	5005	79607	6394	10353	34	3	1	0	0
55	729417	14190	163076	9940	31668	155	8	1	0	0
56	627503	10945	140605	8766	24516	160	3	0	0	0
57	1352876	15576	289910	13154	41446	183	1	0	0	0
58	387318	4178	80299	9980	11211	514	3	0	0	0
59	416400	6446	89453	8376	13739	239	1	1	2	0
60	361595	5567	70789	9124	12774	26	2	0	0	0

Table 1: Corpus Data