UTRECHT UNIVERSITY

Department of Information and Computing Sciences

Applied Data Science master thesis

Exploring the Intercorrelations of Big Five Personality Traits: Comparing Questionnaire-Based Methods and Automated Personality Assessment using BERT and RNN Models

First examiner:

Anastasia Giachanou

Candidate:

Yucheng Chen

Second examiner:

Qixiang Fang

July 2, 2023

Abstract

This study investigates the performance of two deep learning models, RoBERTa and Bi-LSTM, in predicting the Big Five personality traits and capturing the correlations among personality traits from text data. The models' performance was evaluated on two datasets, PAN 2015 and PANDORA, using RMSE and R^2 values. The study found that RoBERTa outperformed Bi-LSTM in predicting personality traits on both datasets. However, both models demonstrated varying performances across the two datasets, highlighting the influence of data diversity on model performance. The study also examined the correlations among predicted personality traits and found that the models could capture the sign of the correlations present in the original datasets. However, divergences in the direction of correlations were observed in instances of weak correlations. Besides, the correlations of original datasets and predictions may not align with the findings in psychological studies, which address the importance of annotations when researching the correlations of predicted personality traits.

Contents

1. Introduction

1.1 Background

According to Sanchez-Roige et al. (2018), personality refers to the stable and consistent patterns of an individual's thoughts, feelings, and behaviors that showcase their unique tendencies and qualities. As an integral component of human psychology, personality has been extensively studied in relation to various aspects of individual and societal functioning. For example, studies by Li et al. (2022), Sanchez-Roige et al. (2018), Shen et al. (2020), Esterwood and Robert (2020), and Yang and Huang (2019) have demonstrated the practical applications of personality across various domains, including psychology, mental health care, personalized marketing strategies, and the development of recommender systems.

In order to understand and measure personality, several taxonomies have been developed, such as the Big Five Factor Personality Model (Goldberg, 1981; McCrae & Costa, 1987), the Myers-Briggs Type Indicator (MBTI) (Myers, 1962) and others. Among these taxonomies, the Big Five has become the most widely accepted and commonly used framework for assessing personality, as it has consistently demonstrated empirical solid support and cross-cultural applicability (Allik & McCrae, 2002; John et al., 2008). The Big Five model comprises five dimensions, which are openness (OPE), conscientiousness (CON), extraversion (EXT), agreeableness (AGR), and neuroticism (NEU).

Conventionally, studies on personality traits have primarily relied on questionnaire-based methods, which depend on self-report or informant-report measures (Fang et al., 2022). Renowned questionnaires such as the Revised NEO Personality Inventory (NEO-PI-R) (Costa Jr & McCrae, 2008) and the Big Five Inventory (BFI) (John et al., 1991) have been extensively used to assess the Big Five personality traits. These questionnaires typically consist of a series of items or statements that participants rate based on their level of agreement or endorsement using a Likert-type scale. The responses are then scored and analyzed to derive personality trait scores or profiles for each individual (John & Srivastava, 1999). Despite their widespread use, Fang et al. (2022) have highlighted that these questionnaire-based personality assessments can be time-consuming and labor-intensive. Moreover, ensuring data availability presents a challenge due to potential participant concerns, such as privacy issues, which may deter individuals from participating in these research studies and affect the volume and quality of collected data. This has prompted the exploration of automated personality assessment methods that leverage user-generated data, giving rise to the field of personality computing (PC). PC is a research area that uses computational techniques, such as natural language processing and machine learning algorithms, to automatically assess and predict personality traits from user-generated data (Phan & Rauthmann, 2021).

Building on this foundation, the advent of machine learning and natural language processing has enabled the development of innovative algorithms and models within the PC field. These advancements facilitate the prediction of personality traits from diverse user-generated data, including text, audio, and images (Ahmad et al., 2021; Guntuku et al., 2020; Sun et al., 2018). Moreover, transformer-based models like Bidirectional Encoder Representations from Transformers (BERT), which are trained based on a large volume of text data, have achieved outstanding performance in various tasks, such as text annotation and sentiment analysis (Hoang et al., 2019; Miller, 2019), and have also accelerated the development of automatic PC.

In previous research, numerous studies have leveraged architectures such as BERT and Recurrent Neural Networks (RNN) in the field of PC. However, despite these advancements, there remains a gap in the literature. According to Fang et al. (2022), almost no studies have reported the intercorrelations among the predicted personality traits in the field of PC, which is crucial for better understanding the underlying structure of personality and ensuring that automatic assessments align with established psychological findings.

Given that the intercorrelations between the factors of the Big Five traits can provide a deeper understanding of human personality and its implications for various aspects of life, several studies have focused on the individual factors of the Big Five personality model and investigated the intercorrelations among these factors. For instance, van der Linden et al. (2010) found significant positive correlations between openness, extraversion, and conscientiousness. DeYoung et al. (2007) reported that neuroticism was negatively correlated with conscientiousness and extraversion. These studies highlight the complex interplay among the Big Five traits and underscore the importance of considering these intercorrelations when assessing personality.

Building on this foundation, this paper aims to investigate and compare the correlations among the predicted Big Five personality traits using personality computing methods, particularly BERT, specifically the Robustly Optimized BERT Pretraining Approach (RoBERTa), and RNN, specifically Bi-directional Long Short-Term Memory (Bi-LSTM). This paper aims to examine whether these techniques, specifically RoBERTa and Bi-LSTM, can not only predict the personality traits of individuals but also accurately represent the intercorrelations among these traits. To fulfill this aim, the following research question and two sub-questions are proposed:

Research Question: How do the correlations among predicted Big Five personality traits in BERT-based models compare to those in RNN-based models and questionnaire-based methods?

Sub-question 1: To what extent can BERT-based, RNN-based models accurately predict user personality traits based on user-generated texts?

Sub-question 2: How do the intercorrelations among the five factors differ when comparing BERT-based models, RNN-based models, and questionnaire-based methods?

2. Literature Review

This section provides a review of prior research pertaining to personality computations. Initially, it explains the concept of personality traits and the prevailing personality taxonomies. Next, the focus shifts to various methodologies employed in assessing an individual's personality. This includes questionnaire-based approaches, and computational techniques utilised in personality detection. Finallyelucidart, the intricate interrelationships among personality traits, as evidenced in prior studies, will be discussed.

2.1 Personality Traits and Big Five

A personality trait assessment serves as a standardized tool for gauging an individual's cognitive, emotional, and behavioral predispositions (Kreuter et al., 2022). Typically, personality trait models establish various classification dimensions and use questionnaires to assess these dimensions (John et al., 2010; Matthews et al., 2003). Among those models, the Big Five and MBTI are the most influential and widely accepted personality taxonomies. However, according to Fang et al. (2022), the Big Five personality taxonomy may offer several advantages over MBTI. First, the Big Five may encompass a more accurate and realistic personality assessment, as it scores individuals along a continuous spectrum and includes facets for finer-grained analysis. Additionally, the Big Five has a more robust empirical foundation based on large-scale quantitative analyses. Its natural language roots suggest Big Five-related cues are likely more prevalent in text data than MBTI-related cues (Fang et al., 2022).

The Big Five model delineates five broad dimensions that encapsulate the core facets of an individual's personality (Goldberg, 1981; McCrae & Costa, 1987):

- Openness to Experience: Characterizing an individual's receptiveness to novel ideas, experiences, and intellectual exploration, this dimension describes those who are highly imaginative, curious, and open-minded (McCrae & Sutin, 2009).

- Conscientiousness: This dimension signifies an individual's level of organization, dependability, and responsibility (Barrick & Mount, 1991).
- Extraversion: Extraversion denotes an individual's assertiveness, sociability, and tendency to display positivity in social situations (Lucas & Baird, 2004).
- Agreeableness: Reflecting an individual's tendency towards cooperation and kindness towards others (Graziano & Eisenberg, 1997).
- Neuroticism: Representing emotional instability, this dimension encapsulates traits such as anxiety, moodiness, and the proneness to experience negative emotions (Lahey, 2009).

These dimensions are not standalone aspects of personality but are correlated and shape an individual's complete personality profile together. Moreover, these personality dimensions have significant implications for predicting individual behaviors and life outcomes. For example, conscientiousness has been found to be a consistent predictor of academic and job performance (Barrick & Mount, 1991), while extraversion is often linked to leadership tendencies and is associated with increased well-being and happiness (Lucas & Baird, 2004).Having established a foundational understanding of the Big Five personality taxonomy, it becomes imperative to delve into the methodologies employed for its assessment.

2.2 Questionnaire-Based Personality Assessment Methods

Methodologies in personality assessment can be broadly categorized into questionnaire-based and computational approaches (Štajner & Yenikent, 2020). The questionnaire-based methods, which are predominantly used in the field of psychology, are grounded in self-reporting techniques. In these methods, participants provide personal responses to a series of questions. The development of these questionnaire instruments involves rigorous validation steps, ensuring a robust empirical foundation that attests to the reliability of these methods.

While there are numerous tools available for personality assessment, a few notable ones include the Ten-Item Personality Inventory (TIPI) by Gosling et al. (2003), Big Five Aspect Scales (BFAS) (DeYoung et al., 2007), and the Revised NEO Personality Inventory (NEO-PI-R). Each of these instruments has its unique strengths and applications in personality assessment. The TIPI, for instance, is a compact tool that swiftly captures a snapshot of an individual's personality, making it particularly useful in time-sensitive situations (Gosling et al., 2003). On the other hand, the BFAS delves deeper, offering a more comprehensive assessment of the Big Five personality traits (DeYoung et al., 2007). This tool is ideal for situations where a more detailed understanding of an individual's personality is required. The NEO-PI-R evaluates the six facets of each Big Five personality trait, providing a thorough and detailed personality profile (Costa Jr & McCrae, 2008). This instrument is often employed in clinical settings or in-depth research studies where a comprehensive understanding of an individual's personality is crucial (Samuel & Widiger, 2008).

While questionnaire-based personality assessment methods are reliable and validated, they do have limitations. They rely heavily on the honesty and self-awareness of the respondent and may not fully capture the complexity of an individual's personality. Furthermore, these methods can be influenced by the respondent's current mood or state of mind (Paulhus & Vazire, 2007). Another potential pitfall is the susceptibility to social desirability bias, where respondents may tailor their answers to align with what they perceive to be socially acceptable or favorable (Holtgraves, 2004). Additionally, recruiting participants and accessing data from a large population can be challenging, potentially limiting the sample size. Moreover, these methods necessitate human involvement in the assessment process, which requires training of the assessors to ensure consistency and accuracy in interpreting the responses.

2.3 Automatic Personality Computing Based on UGTs

In response to the limitations of questionnaire-based methods, personality researchers have turned their attention to an alternative approach: personality prediction, which offers the potential for more implicit measurements of personality traits (Stachl, Au, et al., 2020) and a higher data availability with larger sample size due to the abundance of online user-generated text.

Personality Computing (PC) is an emerging field that combines personality psychology and computer science. It aims to extract personality traits, such as Big Five levels, from machine-sensed data like written texts, digital footprints, and speech patterns using machine-learning approaches (Phan & Rauthmann, 2021). There are mainly two types of automatic personality trait detection via texts: the lexical and open vocabulary machine learning methods (Ren et al., 2021).

For the first method, the most representative approach is the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001). LIWC is designed to analyze text for style and personality-related vocabulary, capturing elements such as the focus of attention, emotional expression, social relationships, cognitive styles, and individual differences. This linguistic analysis could provide a rich dataset for personality prediction. Numerous studies have leveraged the psychological attributes extracted by LIWC as input for machine learning models (Hall & Caton, 2017; Li et al., 2022). Adi et al. (Adi et al., 2018) highlight the utility of LIWC in analyzing user-generated texts for predicting users' personalities from a diverse range of social media platforms. However, one important limitation is that the language support of LIWC is limited, as its dictionary comprises only a select number of languages (Adi et al., 2018). LIWC operates with a predefined dictionary that concentrates solely on statistical-based lexical features, thereby failing to capture the implicit semantic information within a sentence (Li et al., 2022).

The open vocabulary method, on the other hand, leverages machine learning techniques and natural language processing (NLP) to analyze text data without relying on a predefined dictionary. Instead, it identifies patterns and relationships between words and phrases in the text (Schwartz et al., 2013). This approach has been found to be effective in predicting personality traits from user-generated texts on social media platforms (Park et al., 2015).

In recent years, deep learning neural networks and large language models have significantly advanced NLP applications like sentiment analysis and opinion mining (Xue et al., 2018). Various researchers have explored their applicability to personality prediction as well. For instance, Yu and Markov (Yu & Markov, 2017) leveraged deep learning methodologies to predict the personality traits of Facebook users. Similarly, Tandera et al. (Tandera et al., 2017) employed a variety of techniques, including multi-layer perceptron, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and 1-Dimensional Convolutional Neural Networks (1-DCNN) to predict Facebook users' personality based on the Big Five personality attributes.

Moreover, Xue et al. (Xue et al., 2018) proposed an innovative Attention Convolutional Neural Network (AttCNN) model. This model was designed to extract deep semantic features from users' post text, which were then combined with statistical linguistic features. The amalgamation of these features was then fed into a regression algorithm to predict personality based on the Big Five personality attributes.

In a separate study, Arijanto et al. (Arijanto et al., 2021) utilized BERT to predict personality traits from English Twitter datasets. BERT is a large language model that learns the context of a word based on all its surroundings in a text, thus enabling a deeper understanding of language (Devlin et al., 2018). Large language models, such as BERT, are AI models that can understand and generate text similar to human writing. They are trained on various internet texts and can produce creative and nuanced outputs due to their extensive training data (Brown et al., 2020).

The variety of studies and methodologies highlights the potential and adaptability of the open vocabulary method for automatic personality computing.

2.4 Intercorrelations Among Big Five Personality Traits

The interrelationships among the Big Five personality traits have long captivated the interest of psychologists. Questionnaire-based methodologies have provided compelling evidence of the interconnectedness of these personality traits. For instance, Digman (Digman, 1997) assembled 14 studies of inter-scale correlations in the Big Five personality traits, and the mean inter-scale correlation was 0.26. Moreover, a meta-analysis by Rushton and Irwing (Rushton & Irwing, 2008) also suggests significant intercorrelations among the Big Five traits. For instance, the correlation between Openness and Conscientiousness was found to be 0.208, suggesting a modest positive relationship. Similarly, a stronger positive correlation of 0.413 was observed between Extraversion and Openness, indicating that outgoing and sociable individuals (EXT) often also tend to be open to new experiences (OPE). The study also found positive correlations between Agreeableness and both Openness (0.114) and Conscientiousness (0.413).

Building on these findings, Van der Linden et al. (van der Linden et al., 2010) also studied the intercorrelations among the Big Five traits. The study highlighted a corrected correlation of 0.43 between Openness and Extraversion; this suggests a positive relationship between these two traits. Additionally, the corrected correlation between Conscientiousness and Agreeableness is 0.43, while the intercorrelation between Conscientiousness and Neuroticism is -0.43. Furthermore, the study examined the intercorrelations across various questionnaires and sample populations, reinforcing the consistency of these correlations across different questionnaires and populations.

Based on the above studies, it is commonly considered that certain pairs of the Big Five personality traits often correlate with each other. For instance, individuals who are open to new experiences (OPE) often also exhibit outgoing and sociable behaviors (EXT). Similarly, those who are organized, dependable, and disciplined (CON) often also display cooperative, warm, and considerate behaviors (AGR). Furthermore, outgoing and energetic individuals (EXT) often also tend to be friendly, empathetic, and get along well with others (AGR).

On the other hand, there are also pairs of traits that often show negative correlations in various studies. For example, individuals who tend to experience negative emotions like anxiety and anger (NEU) often have lower scores in Conscientiousness, indicating potential difficulties with organization and discipline. Similarly, those who score high on Neuroticism often score low on Agreeableness, suggesting potential struggles with cooperation and consideration of others. Lastly, individuals with high Neuroticism often score low on Extraversion, indicating that they may be less outgoing and energetic and more reserved or withdrawn (Rushton & Irwing, 2008; van der Linden et al., 2010).

While these correlations are often seen in research, they should not be interpreted as fixed or absolute. Additionally, in the field of automatic personality prediction, the examination of intercorrelations among traits is not as frequent. Considering these intercorrelations is a crucial part of personality trait theory, and their inclusion in automatic personality prediction could offer a deeper understanding of the complex interactions among predicted personality traits.

3. Method

This section outlines our methodology, including several stages: data collection, preprocessing, model development, evaluation, and comparison. Initially, this section will introduce the two datasets used in this research, followed by a detailed description of the data collection process. The next stage involves outlining the preprocessing steps for the datasets. Following this, the architecture of the RoBERTa and Bi-LSTM models will be presented, focusing on each model's components, configurations, and training procedures.

3.1 Data Collection

Acquiring personality data is a complex task due to privacy concerns and the financial burden of hiring professional psychologists for accurate labeling (Ren et al., 2021). As a result, utilizing publicly available personality datasets is a common practice in the Personality Computing field.

In this study, two datasets are employed: the Author Profiling dataset from PAN@CLEF 2015 (Rangel et al., 2015) and the PANDORA dataset (Gjurković et al., 2020). The Author Profiling dataset comprises 294 users along with their tweets, demographic information, and self-reported Big Five personality dimensions: Openness (OPN), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU). Access to this dataset was requested and granted by the PAN@CLEF 2015 organizers, ensuring that the data was used in compliance with their data usage policy and ethical guidelines.

The PANDORA dataset mainly contains two files, with the first file including over 17 million Reddit comments from around 10,000 users and another file recording the user profiles, while the personality traits are collected from various self-assessed tests.

In this study, the Author Profiling dataset from PAN@CLEF 2015 and the PANDORA dataset were chosen due to their availability, relevance, size, diversity, and ethical considerations. Both datasets are publicly accessible and contain text data along with personality labels, making them highly suitable for personality prediction tasks. The PANDORA dataset, with over 17 million Reddit comments, offers a large and diverse set of data, while the PAN@CLEF 2015 dataset provides complementary data from Twitter. Importantly, these datasets have been collected and shared in accordance with ethical guidelines, addressing potential privacy concerns. Therefore, their use not only aligns with the research objectives but also adheres to ethical standards in data usage (Gjurković et al., 2020; Rangel et al., 2015).

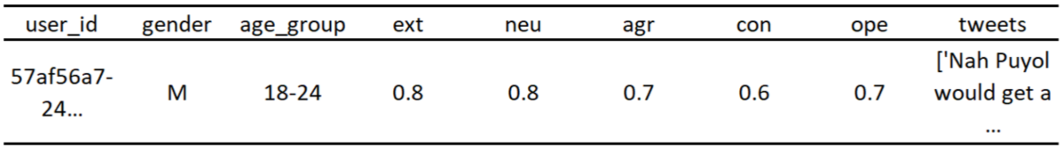
3.1.1 PAN 2015 Author Profiling Dataset

The PAN 2015 Author Profiling dataset, available in various languages, provides both training and testing data. This dataset comprises multiple XML files, each containing user-generated tweets and named according to user IDs. Alongside these, a "truth.txt" file provides users' demographic data.

The data extraction process begins with the collection of data from the XML files. Subsequently, demographic and personality information is retrieved from the "truth.txt" file using the user ID that corresponds to the XML file names. This information, including user IDs, personality traits, and tweets, is then appended to designated lists and exported to a single CSV file for training and testing data.

Originally, the dataset contained 294 entries, each corresponding to a unique user. In the initial format, each entry in the 'tweets' column contained a list of tweets from the respective user. However, for the purpose of this study, the data has been restructured such that each tweet is represented as a separate line, thereby expanding the dataset to include one line per tweet. The structure of the restructured dataset is visually represented in Table 1.

It is important to note that, as the comparison will be conducted between two datasets, the scale of trait values has been normalised to a range of 0-1 to ensure consistency. In both datasets, Min-Max Scaling has been applied to normalise the values for personality traits. This involves using a range of (-0.5, 0.5) for the PAN 2015 dataset and (0, 100) for the PANDOR dataset in accordance with the formula specified below.



user_id	gender	age_group	ext	neu	agr	con	ope	tweets
57af56a7-24...	M	18-24	0.8	0.8	0.7	0.6	0.7	['Nah Puyol would get a ...

Table 1: PAN 2015 Author Profiling Dataset

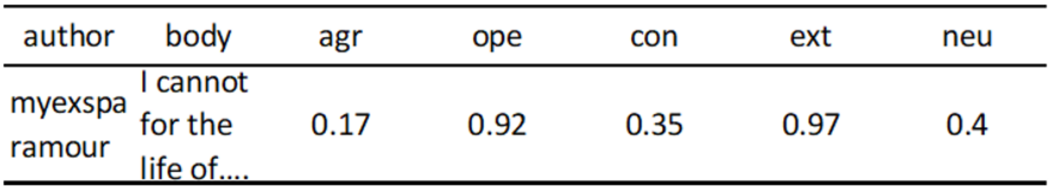
$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Figure 1: Min-Max Scaling formula

3.1.2 PANDORA dataset

The PANDORA dataset comprises two CSV files: one that stores user-generated texts and another that contains author profiles. However, many authors in the dataset have null values for personality traits. For the purpose of this research, only users with non-null personality values are included, resulting in a total of 1568 users.

Following this, comments from these selected authors are extracted and merged with the corresponding author profiles. This process yields a total of 3,006,566 comments. Given the substantial size of the PANDORA dataset, a random sample of 20,000 comments from 134 authors has been selected for the purpose of this study. The structure of the sampled dataset, after the selection and merging processes, is visually represented in Table 2.



author	body	agr	ope	con	ext	neu
myexspa ramour	I cannot for the life of....	0.17	0.92	0.35	0.97	0.4

Table 2: PANDORA Dataset

3.2 Data Cleaning

Prior to analyzing the collected data and developing the prediction models, it is crucial to preprocess the user-generated text. Text cleaning is a standard practice in natural language processing and machine learning, as it can significantly improve the performance of models by reducing noise and focusing on relevant features (Haddi et al., 2013).

Notably, the pre-processed texts will only be used for the RNN-based model, specifically the Bi-LSTM, as RNNs can benefit from such preprocessing steps. BERT, on the other hand, handles the understanding of the meaning of a word in its context (polysemy) and does not require lemmatization or stemming. Moreover, BERT has its own method of handling stopwords in its subword tokenization process. Therefore, preprocessing steps like removing stopwords, lemmatization, and converting to lowercase are not required when using BERT (Devlin et al., 2018). Moreover, BERT is capable of understanding the meaning of words in different contexts, making it more robust to variations in the language (Devlin et al., 2018).

The text preprocessing begins with the raw text input. Firstly, slang and abbreviations are replaced with their standard forms to ensure the text is understandable and consistent. Following this, any URLs and mentions (identified by ’@username’) are removed from the text. The next phase involves converting the text to lowercase, which aids in maintaining uniformity across the text and prevents word duplication due to case variations. Following this, the text is stripped of punctuation and stopwords from the NLTK library (Bird et al., 2009), including custom stopwords like "hahaha," "ahhh," "rt," and others to reduce the dimensionality of word representations. This pivotal step aids in reducing the data's dimensionality and shifts the focus onto the text's important words. Finally, the text is lemmatized, which involves reducing words to their base or root form (for example, 'running' to 'run'). This helps in grouping together different inflections of a word and treating them as a single item. Figure 1 shows the steps of data cleaning.

3.3 Data Preprocessing

After cleaning the data, the next step is data preprocessing, which transforms the data into recognisable input for the models. This process can be various depending on the specific requirements of the models being used.

For the RNN-based model (Bi-LSTM), the preprocessed text is tokenised. Tokenisation is the process of splitting the text into individual words or "tokens". This is crucial as it converts the text data into a format the model can understand. In this case, the preprocessed text is initially tokenised and encoded into integers as sequences; then, the sequences are padded or truncated to ensure they all have the same length.

The preprocessing steps for BERT-based models, such as RoBERTa, differ from traditional models. BERT uses its own tokeniser, which uses WordPiece tokenisation to break words into subwords. These subwords are then mapped to a vocabulary of known words. The advantage of this approach is that it can handle words not present in the vocabulary by breaking them down into subwords (Ma et al., 2020). It also adds special tokens like "[CLS]" and "[SEP]" to indicate the start and end of sentences. Unlike some tokenisers that consider words individually, the BERT tokeniser and the BERT model consider the context of words within a sentence (Devlin et al., 2018).

For both models, the processing step requires setting a max length of padding to ensure all texts have the same length. For Bi-LSTM, the max length is set as 128 for both datasets, while the maximum length for RoBERTa is set to the longest text in the dataset in this case since RoBERTa could handle long texts better than Bi-LSTM.

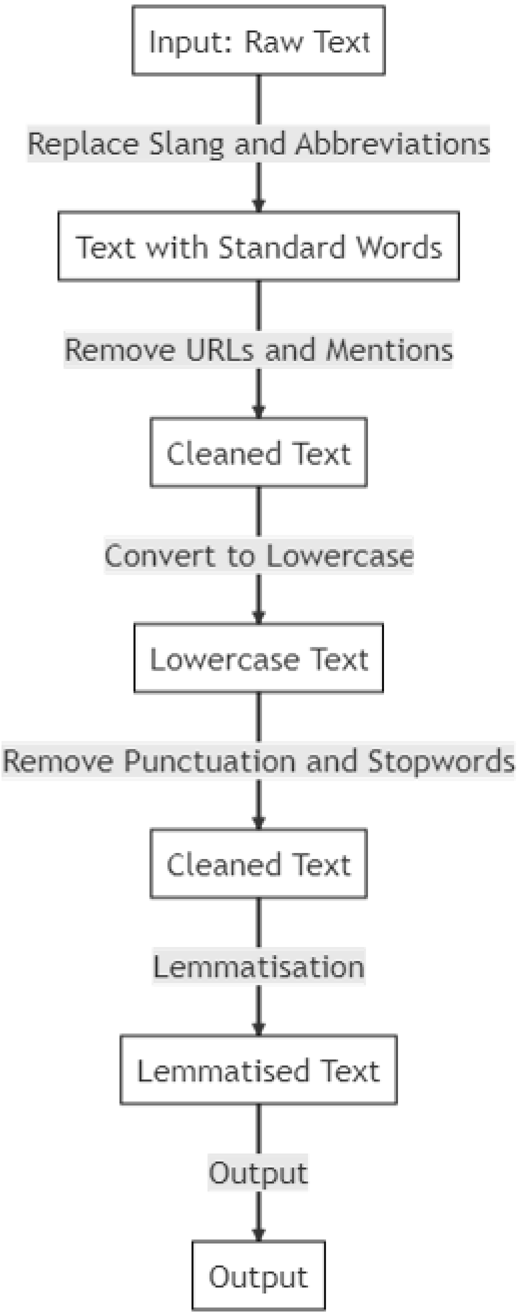


Figure 2: Workflow of Data Cleaning

3.4 Model Development

The model development stage involves the design and training of the two models used in this study: RoBERTa and Bi-LSTM. This section will outline the architecture of each model, the configuration settings, and the training procedures.

3.4.1 RoBERTa Model

RoBERTa is a variant of BERT that uses a different training approach and has been shown to outperform the original BERT model on several benchmarks (Liu et al., 2019). Despite sharing the same transformer-based architecture with BERT, which incorporates several layers of self-attention mechanisms, RoBERTa distinguishes itself through its pre-training corpus, which allows it to learn the underlying structure of the language and the context of words. This pre-trained model is then fine-tuned for the specific task of personality prediction.

The RoBERTa model used in this study is a pre-trained model provided by the Hugging Face library (Liu et al., 2019; Roberta-base, n.d.). The input to the RoBERTa model is the raw text data from the datasets. The model processes this raw text and outputs a vector representation for each input text. The vector representation is then passed through a dropout layer, which randomly sets a fraction of input units to 0 at each update during training time to prevent overfitting. The dropout rate, in this case, is set to 0.3.

Following the dropout layer, the vector representation is passed through a fully connected linear layer. This linear layer maps the high-dimensional input vector to a lower dimension corresponding to the Big Five personality traits. The output size of this linear layer is set to 5 to corporate with a multi-task regression problem corresponding to the five personality traits.

The model is trained using the Adam optimiser with a learning rate of 1e-5. The loss function used is the Mean Squared Error (MSE), which is suitable for regression tasks like personality prediction. The model is trained for 10 epochs. Figure 2 shows the structure of the RoBERTa model used in

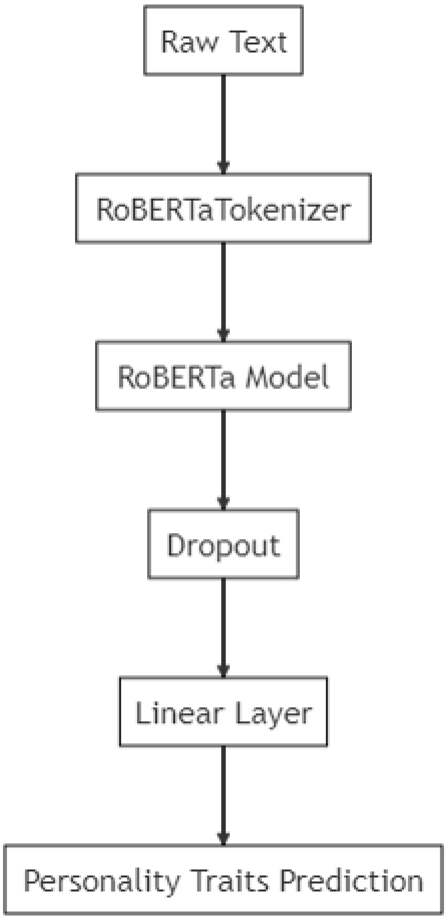


Figure 3: Structure of RoBERTa

this study. Fine-tuning was not applied when choosing the dropout rate and the number of epochs, and it was only applied to the dataset PAN 2015, Due to the time-consuming nature of fine-tuning. This decision was made to balance the need for model optimisation with the practical constraints of computational resources and time.

3.4.2 Bi-LSTM Model

The Bi-LSTM model is a type of RNN that is capable of learning long-term dependencies in text data (Hochreiter & Schmidhuber, 1997). The bidirectional aspect of the model allows it to learn the context of a word based on the words that come before and after it (Schuster & Paliwal, 1997).

The architecture of the Bi-LSTM model used in this study consists of an embedding layer, a Bi-LSTM layer, and a fully connected layer. The em-

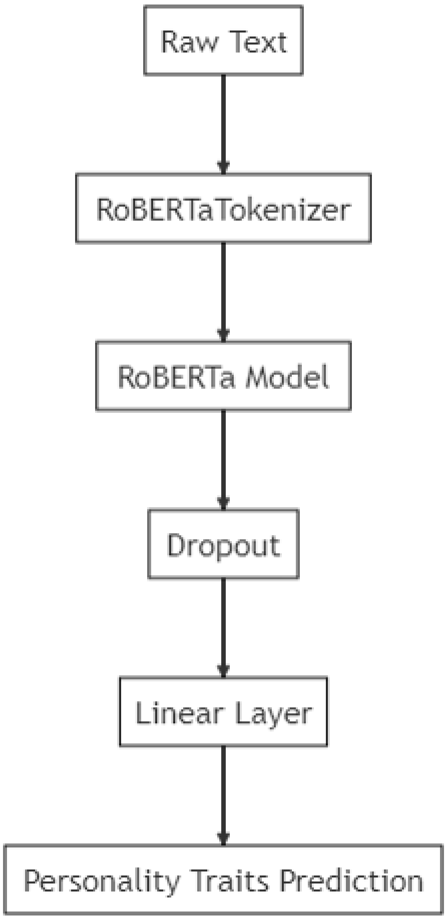
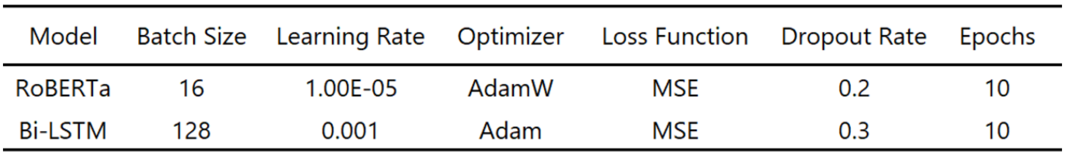


Figure 4: Structure of RoBERTa

bedding layer transforms the integer-encoded text data into dense vector representations. The Bi-LSTM layer processes these vector representations and outputs another set of vectors that capture the contextual information in the text. The fully connected layer takes these vectors and predicts the Big Five personality traits. Figure 3 illustrates the structure of the adopted Bi-LSTM model.

In the process of training both RoBERTa and Bi-LSTM models, specific hyperparameters are set. RoBERTa's batch size is 16, and the learning rate is 1e-5. On the other hand, the Bi-LSTM model is configured with a batch size of 128 and a learning rate of 0.001. The choices of these hyperparameters ared conducted via a grid search strategy. However, due to the time-consuming nature of the training process, the grid search method for the RoBERTa model is only applied to the selection of the batch size, with options being 16, 32, 64, and 128.

For the Bi-LSTM model, the grid search strategy is applied to both the batch size and the learning rate. The batch size options remain the same as those for RoBERTa, while the search range for learning rate consisted of the following values: 0.1, 0.01, 0.001, 0.0001, and 0.00001. The settings of the models are listed in Table 3 below.



Model	Batch Size	Learning Rate	Optimizer	Loss Function	Dropout Rate	Epochs
RoBERTa	16	1.00E-05	AdamW	MSE	0.2	10
Bi-LSTM	128	0.001	Adam	MSE	0.3	10

Table 3: Hyperparameter Settings

4. Results

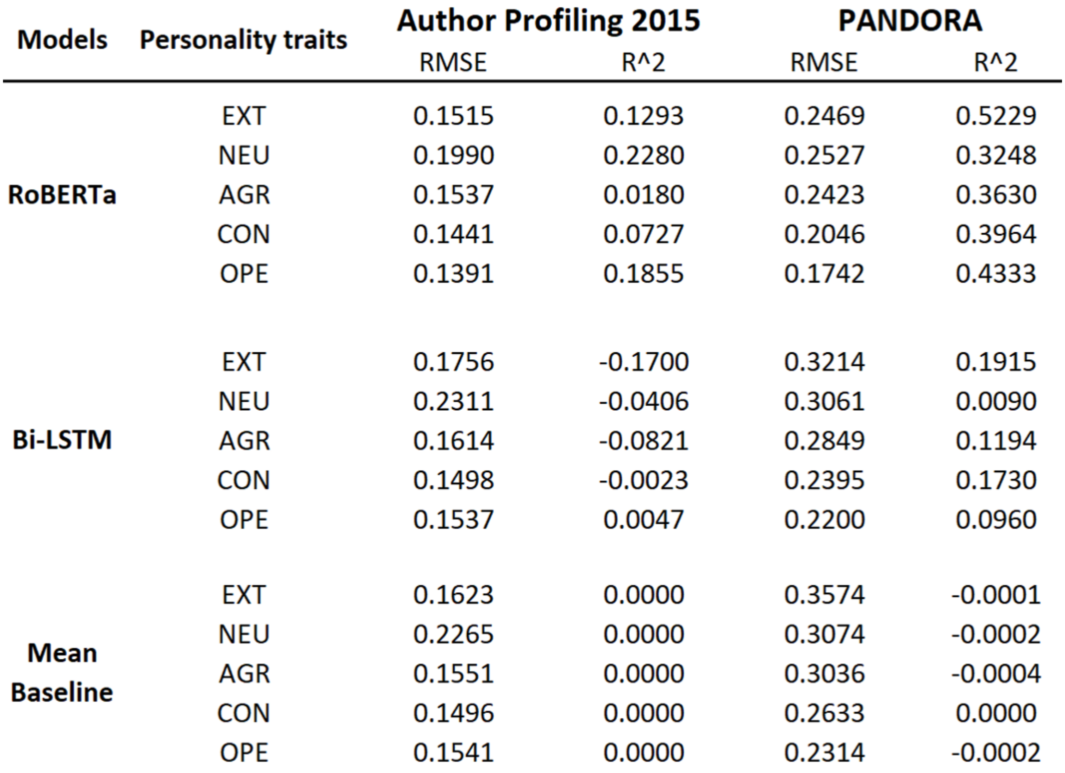
This section presents the findings of the study. It begins with the evaluation of the three models, including the baseline model across two datasets. It includes comparing the models' performance metrics by their Root Mean Square Error (RMSE) and R^2. Finally, the section explores the correlations among personality traits as predicted by the models in both datasets.

4.1 Model Evaluation and Comparison

The performance metrics used for comparison are RMSE and R^2. RMSE and R^2 were chosen as performance metrics because they offer complementary perspectives on model performance. RMSE measures the average prediction error, providing a sense of how much, on average, the predictions deviate from the actual values. On the other hand, R^2, also known as the coefficient of determination, provides a measure of how well the model's predictions fit the actual values in terms of explained variance. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variable, thus providing an understanding of the goodness of fit of the model.

Table 4 below illustrates the performance metrics of all three models on the two datasets for all personality traits. Lower RMSE values suggest better model performance, the R^2 values, on the other hand, represent the proportion of the variance in the personality traits that is predictable from the user-generated texts, with higher values indicating a better fit of the model to the data.

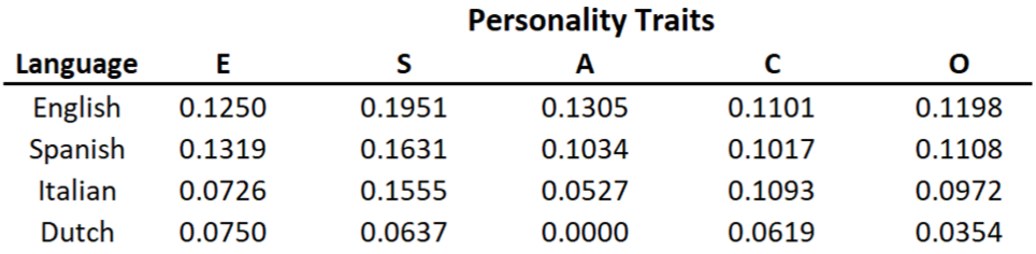
For the PAN 2015 Author Profiling dataset, we notice that RoBERTa achieved lower RMSE values across all personality traits compared to the Bi-LSTM and mean baseline. Specifically, for the RoBERTa model, the RMSE values range from 0.1391 (OPE) to 0.1990 (NEU). In contrast, for Bi-LSTM,



Models	Personality traits	Author Profiling 2015		PANDORA	
		RMSE	R^2	RMSE	R^2
RoBERTa	EXT	0.1515	0.1293	0.2469	0.5229
	NEU	0.1990	0.2280	0.2527	0.3248
	AGR	0.1537	0.0180	0.2423	0.3630
	CON	0.1441	0.0727	0.2046	0.3964
	OPE	0.1391	0.1855	0.1742	0.4333
Bi-LSTM	EXT	0.1756	-0.1700	0.3214	0.1915
	NEU	0.2311	-0.0406	0.3061	0.0090
	AGR	0.1614	-0.0821	0.2849	0.1194
	CON	0.1498	-0.0023	0.2395	0.1730
	OPE	0.1537	0.0047	0.2200	0.0960
Mean Baseline	EXT	0.1623	0.0000	0.3574	-0.0001
	NEU	0.2265	0.0000	0.3074	-0.0002
	AGR	0.1551	0.0000	0.3036	-0.0004
	CON	0.1496	0.0000	0.2633	0.0000
	OPE	0.1541	0.0000	0.2314	-0.0002

Table 4: Modles Performance

the range is from 0.1498 (CON) to 0.2311 (NEU), and for the baseline model, it is from 0.1496 (CON) to 0.2265 (NEU). On average, the RMSE value of the RoBERTa model is 0.0169 lower than the Bi-LSTM model, with original personality traits values ranging from 0 to 1. Rangel et al. (2015) also indicate the best results in RMSE for predicting personality traits which are shown in Table 5 below. In comparison to these benchmarks, our RMSE values for all personality traits are, on average, 17.6% higher than the best performances of all teams.



	Personality Traits				
Language	E	S	A	C	O
English	0.1250	0.1951	0.1305	0.1101	0.1198
Spanish	0.1319	0.1631	0.1034	0.1017	0.1108
Italian	0.0726	0.1555	0.0527	0.1093	0.0972
Dutch	0.0750	0.0637	0.0000	0.0619	0.0354

Table 5: Best Perfrormance on PAN 2015 dataset

From the perspective of R^2 values, RoBERTa also outperformed Bi-LSTM. The R^2 values for RoBERTa ranged from 0.0180 (AGR) to 0.2280 (NEU), indicating a moderate level of explained variance. In contrast, the R^2 values for Bi-LSTM were negative for most traits, suggesting that the model did not fit the data well.

When applied to the PANDORA dataset, the RoBERTa model again outperformed the Bi-LSTM model. The RMSE values for RoBERTa ranged from 0.1742 (OPE) to 0.2527 (NEU), for Bi-LSTM, they ranged from 0.2200 (OPE) to 0.3214 (EXT), while for the baseline model, it is from 0.2314 (OPE) to 0.3574 (EXT). This suggests that the RoBERTa model was more accurate in predicting personality traits for the PANDORA dataset as well.

The R^2 values for RoBERTa on the PANDORA dataset ranged from 0.3248 (NEU) to 0.5229 (EXT), indicating a substantial level of explained variance. In contrast, the R^2 values for Bi-LSTM were significantly lower, ranging from 0.0090 (NEU) to 0.1915 (EXT). Generally, the R^2 for both models on both datasets are low, however, in the domain of predicting human behaviours, such as personality traits, achieving high R^2 values is challenging, thus, in this case, the R^2 values of RoBERTa in PANDORA dataset could be regarded as relatively high. To my best knowledge, there are no similar studies using R^2 values as performance indicators on both datasets. Thus, it is hard to compare the outcomes of this study with the previous one.

4.2 Correlations among Personality Traits

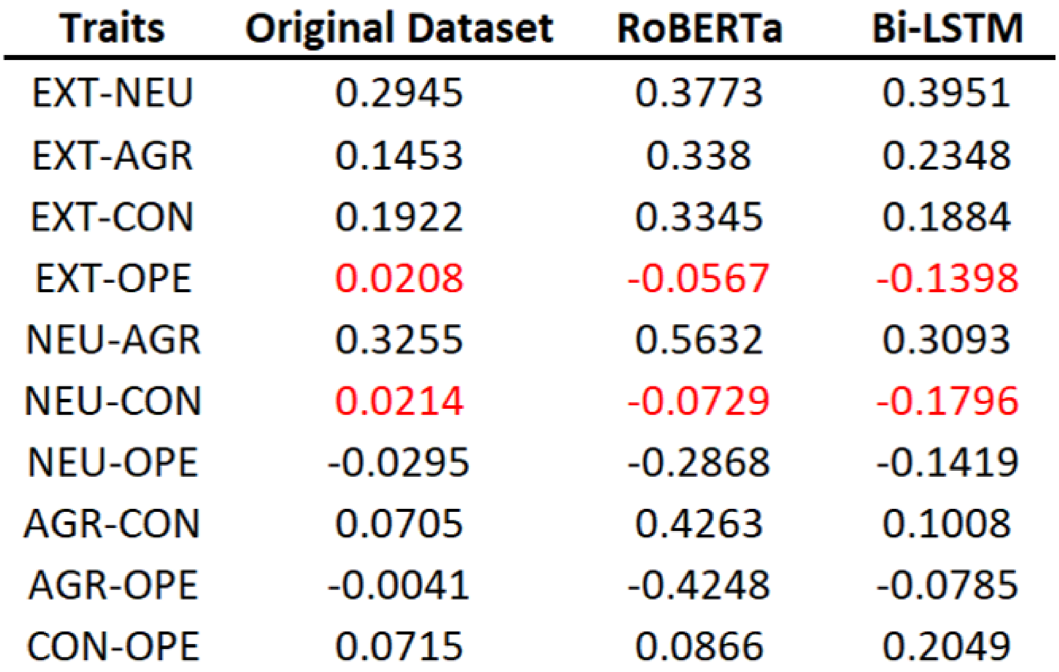
As discussed in the previous section, various studies from the psychological field indicate that there are intercorrelations that exist among personality traits (Rushton & Irwing, 2008; van der Linden et al., 2010). It is crucial to understand those intercorrelations since it may provide insights into how different aspects of personality are related to each other. Table 6 shows the correlations among predicted personality traits in 7039 lines of data from PAN 2015.

Starting with the correlation between Neuroticism (NEU) and Extraversion (EXT), the original dataset showed a correlation of 0.2945. The RoBERTa model predicted a slightly higher correlation of 0.3773, while the Bi-LSTM model predicted an even higher correlation of 0.3951. This suggests that both models could capture the positive relationship between NEU and EXT, although with a slightly stronger correlation than observed in the original dataset.

Next, look at the correlation between Extraversion (EXT) and Agreeableness (AGR). The original dataset showed a correlation of 0.1453. The RoBERTa model predicted a correlation of 0.3380, while the Bi-LSTM model predicted a correlation of 0.2348. Both models predicted a stronger positive correlation between EXT and AGR compared to the original dataset with a higher correlation.

For the correlation between Extraversion (EXT) and Conscientiousness (CON), the original dataset showed a correlation of 0.1922. The RoBERTa model predicted a correlation of 0.3345, while the Bi-LSTM model predicted a correlation of 0.1884. The RoBERTa model predicted a stronger positive correlation between EXT and CON, while the Bi-LSTM model's prediction was closer to the original dataset.

Lastly, for the correlation between Extraversion (EXT) and Openness (OPE), the original dataset showed a correlation of 0.0208. The RoBERTa model predicted a negative correlation of -0.0567, while the Bi-LSTM model predicted a negative correlation of -0.1398. Both models predicted a negative correlation between EXT and OPE, which is a divergence from the positive correlation observed in the original dataset. However, it is important to note that the positive correlation in the original dataset is quite low, suggesting a weak relationship. Similarly, the negative correlations predicted by the RoBERTa and Bi-LSTM models are also low, indicating a weak inverse relationship. This may suggest that when the strength of this relationship is weak, the models may identify a different direction of correlation.



Traits	Original Dataset	RoBERTa	Bi-LSTM
EXT-NEU	0.2945	0.3773	0.3951
EXT-AGR	0.1453	0.338	0.2348
EXT-CON	0.1922	0.3345	0.1884
EXT-OPE	0.0208	-0.0567	-0.1398
NEU-AGR	0.3255	0.5632	0.3093
NEU-CON	0.0214	-0.0729	-0.1796
NEU-OPE	-0.0295	-0.2868	-0.1419
AGR-CON	0.0705	0.4263	0.1008
AGR-OPE	-0.0041	-0.4248	-0.0785
CON-OPE	0.0715	0.0866	0.2049

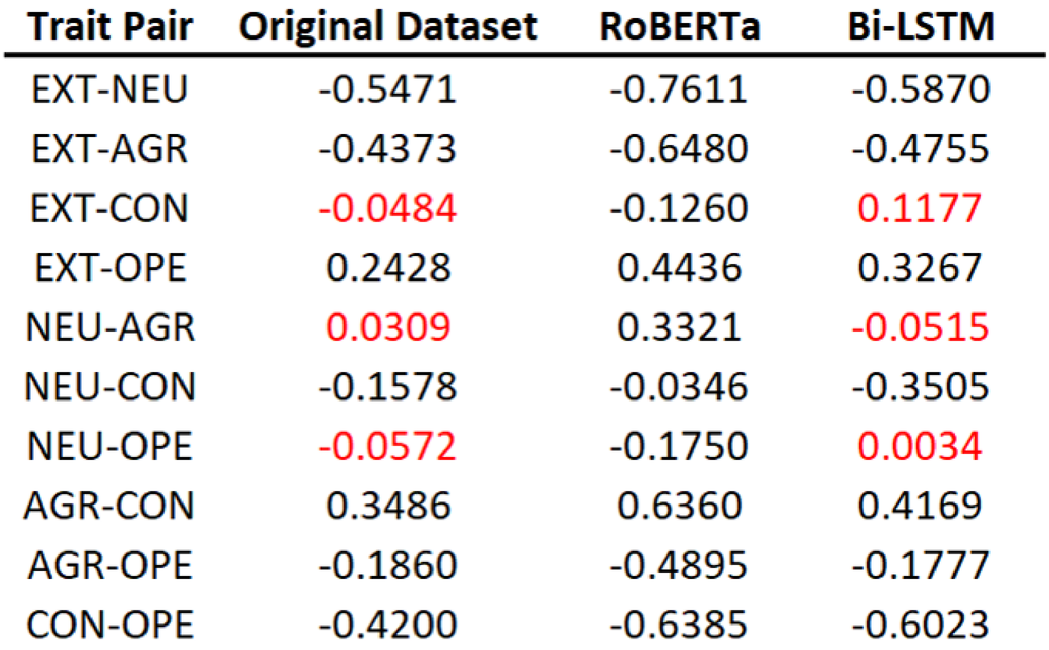
Table 6: Comparative Analysis of Personality Trait Correlations: Original PAN 2015 Dataset vs. Predictions by RoBERTa and Bi-LSTM Models (Note. The red font in the table indicates a change in the direction of correlations between the original dataset and the predicted outputs.)

For the correlations among traits for the PANDORA dataset, we observe similar trends with some variations. For the correlation between NEU and EXT, the sampled dataset showed a correlation of -0.5471. The RoBERTa model predicted a stronger negative correlation of -0.7611, while the Bi-LSTM model predicted a slightly weaker negative correlation of -0.5870. This suggests that both models were able to capture the negative relationship between NEU and EXT, with the RoBERTa model predicting a notably stronger negative correlation.

Next, look at the correlation between EXT and AGR. The sampled dataset showed a negative correlation of -0.4373. The RoBERTa model predicted a stronger negative correlation of -0.6480, while the Bi-LSTM model predicted a slightly weaker negative correlation of -0.4755. Both models predicted a stronger negative correlation between EXT and AGR compared to the sampled dataset, with the RoBERTa model predicting a notably stronger negative correlation.

For the correlation between EXT and CON, the sampled dataset showed a correlation of -0.0484. The RoBERTa model predicted a negative correlation of -0.1260, while the Bi-LSTM model predicted a positive correlation of 0.1177. The RoBERTa model predicted a stronger negative correlation between EXT and CON, while the Bi-LSTM model's prediction diverged from the sampled dataset, indicating a positive correlation.

Lastly, for the correlation between EXT and OPE, the sampled dataset showed a positive correlation of 0.2428. The RoBERTa model predicted a stronger positive correlation of 0.4436, while the Bi-LSTM model predicted a positive correlation of 0.3267. Both models predicted a positive correlation between EXT and OPE, which aligns with the correlation observed in the sampled dataset but with stronger correlations predicted by the models.



Trait Pair	Original Dataset	RoBERTa	Bi-LSTM
EXT-NEU	-0.5471	-0.7611	-0.5870
EXT-AGR	-0.4373	-0.6480	-0.4755
EXT-CON	-0.0484	-0.1260	0.1177
EXT-OPE	0.2428	0.4436	0.3267
NEU-AGR	0.0309	0.3321	-0.0515
NEU-CON	-0.1578	-0.0346	-0.3505
NEU-OPE	-0.0572	-0.1750	0.0034
AGR-CON	0.3486	0.6360	0.4169
AGR-OPE	-0.1860	-0.4895	-0.1777
CON-OPE	-0.4200	-0.6385	-0.6023

Table 7: Comparative Analysis of Personality Trait Correlations: Sampled PANDORA Dataset vs. Predictions by RoBERTa and Bi-LSTM Models (Note. The red font in the table indicates a change in the direction of correlations between the original dataset and the predicted outputs.)

Both the RoBERTa and Bi-LSTM models were able to capture the intercorrelations among the Big Five personality traits to varying degrees. However, there were some discrepancies between the predicted correlations and those observed in the sampled dataset. This suggests that while these models can predict individual personality traits, capturing the complex interplay among these traits remains challenging.

5. Discussion

This section discusses the use of R^2 values in model performance evaluation, performance differences across datasets, correlations among predicted personality traits, and the influence of annotation on these correlations. It also indicates the limitations of our study, including potential improvements in data collection, data preprocessing challenges, and constraints in model training.

5.1 R^2 values in measuring models' performance

RMSE, as an important metric to measure the performance of models, is commonly used as a benchmark for various studies in the field of personality computation. However, the R^2 value, which is not constantly mentioned among the studies related to automatic personality detection, is also a useful metric in measuring the performance of models since it may provide a different perspective of results.

R^2 quantifies the degree to which our model explains the variation in personality traits. An R^2 value of 1 indicates that the model perfectly predicts personality traits, while an R^2 value of 0 indicates that the model does not explain any of the variability of the outcome data around its mean. There are also negative R^2 values which may indicate the risk of overfitting (Stachl, Pargent, et al., 2020).

In this study, the R^2 values of RoBERTa were higher than those for the Bi-LSTM model for all personality traits. This suggests that RoBERTa was able to explain a more significant proportion of the variance in personality traits. However, the R^2 values for RoBERTa were still relatively low, indicating a large variance in personality traits that are not captured. In particular, in the PAN 2015 Author Profiling dataset, the R^2 values for the RoBERTa model ranged from 0.0180 (AGR) to 0.2280 (NEU), while for the Bi-LSTM model, the R^2 values were even lower, ranging from -0.1700 (EXT) to 0.0047 (OPE).

Interestingly, we also observed that the R^2 values varied between the two datasets. For the PAN 2015 dataset, the R^2 values were lower than for the PANDORA dataset. This suggests that the models could better capture the variance in personality traits in the PANDORA dataset. While there could be several factors contributing to the differences observed between the two datasets, such as variations in data quality or quantity, it is likely that one important factor is the diversity of the data in this case.

Furthermore, some traits like EXT and NEU predicted by Bi-LSTM in PAN 2015 demonstrated negative R^2 values. Typically, a negative R^2 suggests that the model does not fit the data well. In our study, this implies that the predictions from the model are less accurate than a simple strategy of using the mean of the personality trait as the prediction. This observation suggests the limitations of using only RMSE as a performance measure.

5.2 Performance Differences across Datasets

The performance of the RoBERTa and Bi-LSTM models also varied significantly across the two datasets used in our study. What may influence the models' ability to predict these traits may be the difference between those two datasets is the distribution of personality traits as figure 5 shows.

The PAN 2015 dataset has a relatively high mean value for all personality traits, with the mean values ranging from 0.6316 (AGR) to 0.7604 (OPE). The standard deviations are relatively low (from 0.1505 to 0.2289), indicating that the personality trait scores are closely clustered around the mean. This suggests that the PAN 2015 dataset may have less variability in the values of personality traits, which could potentially make it easier for the models to predict these traits.

In contrast, the PANDORA dataset has lower mean values for personality traits, ranging from 0.3170 (CON) to 0.7659 (OPE). The standard deviations (from 0.2327 to 0.3585) are higher than the PAN 2015 dataset, indicating a wider spread of personality trait scores. This suggests that the

PANDORA dataset has larger data variability in the values of personality traits, which could make it more challenging for models to predict personality traits.

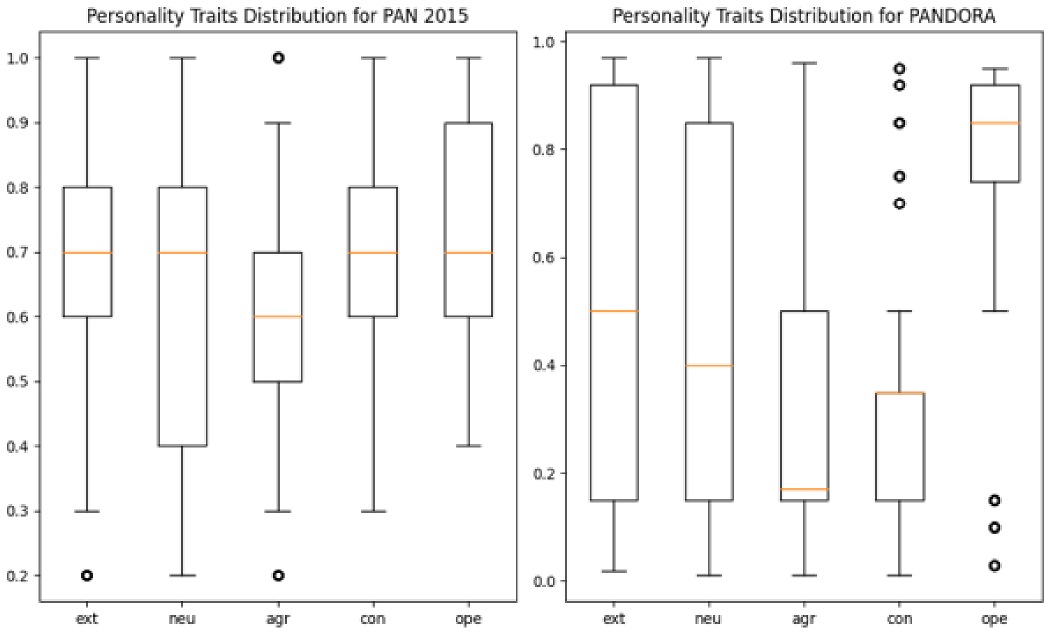


Figure 5: Distribution of Personality Traits values

Additionally, the PANDORA dataset contains fewer authors (134 authors) compared to the PAN 2015 dataset (294 authors). This reduced author diversity could potentially impact the models' performance, contributing to a higher RMSE as the models have less varied author data to learn from.

Interestingly, despite the larger variability in the PANDORA dataset, both models achieved higher R² values with this dataset compared to the PAN 2015 dataset, which suggests that the models were better able to capture the variance in personality traits in the PANDORA dataset. This could be due to the richer context provided by greater diversity in personality trait values in the PANDORA dataset, which may have allowed the models to learn more complex patterns and relationships between the text data and the personality traits.

5.3 Correlations from predicted personality traits

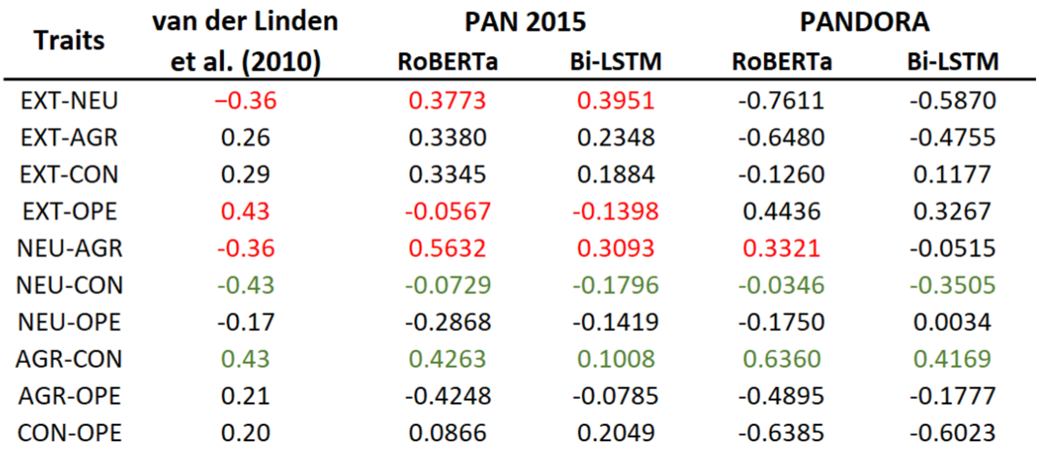
In the analysis of the correlations among predicted personality traits, there are some inconsistencies across different studies and models. The correlation between EXT and OPE found a positive correlation (0.43) in the study by van der Linden et al. (2010). While, in this case, PAN 2015 showed a negative correlation for both the RoBERTa (-0.0567) and Bi-LSTM (-0.1398) models. However, in the PANDORA, the correlation turned positive again for both models, with 0.4436 for RoBERTa and 0.3267 for Bi-LSTM.

The correlation between EXT and NEU was found to be negative in the study by van der Linden et al. (2010) (-0.36). However, in the PAN 2015 study, this correlation was positive for both the RoBERTa (0.3773) and Bi-LSTM (0.3951) models, which is possibly influenced by the correlations of the original PAN 2015 (0.2945). In the PANDORA study, the correlation turned negative again for both models, with -0.7611 for RoBERTa and -0.5870 for Bi-LSTM.

There are also some consistencies across studies and model outputs. For example, the correlation between NEU and CON was found to be negative in the study by van der Linden et al. (2010) (-0.43). In the PAN 2015 study, this correlation was also negative for the RoBERTa model (-0.0729) and Bi-LSTM (-0.1796). In the PANDORA dataset, the correlation was negative for both models, with -0.0346 for RoBERTa and -0.3505 for Bi-LSTM.

In addition, the correlation between AGR and CON was found to be positive in the study by van der Linden et al. (2010) (0.43). This positive correlation was also observed in the PAN 2015 study for the RoBERTa model (0.4263), but was slightly positive for the Bi-LSTM model (0.1008). In the PANDORA study, the correlation was strongly positive for both models, with 0.6360 for RoBERTa and 0.4169 for Bi-LSTM.

Moreover, the correlation between NEU and AGR was found to be negative in the study by van der Linden et al. (2010) (-0.36). In the PAN 2015 study, this correlation was positive for the RoBERTa model (0.5632), but slightly positive for the Bi-LSTM model (0.3093). In the PANDORA study, the correlation was positive for the RoBERTa model (0.3321), but slightly negative for the Bi-LSTM model (-0.0515). Table 8 below shows the correlations from van der Linden et al. (2010)'s studies and the models' predictions.



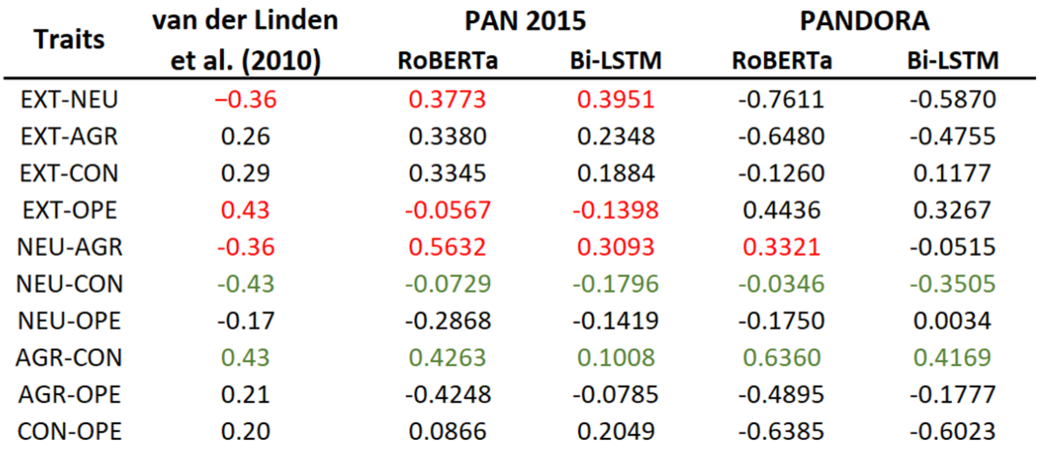
Traits	van der Linden et al. (2010)	PAN 2015		PANDORA	
		RoBERTa	Bi-LSTM	RoBERTa	Bi-LSTM
EXT-NEU	−0.36	0.3773	0.3951	-0.7611	-0.5870
EXT-AGR	0.26	0.3380	0.2348	-0.6480	-0.4755
EXT-CON	0.29	0.3345	0.1884	-0.1260	0.1177
EXT-OPE	0.43	-0.0567	-0.1398	0.4436	0.3267
NEU-AGR	-0.36	0.5632	0.3093	0.3321	-0.0515
NEU-CON	-0.43	-0.0729	-0.1796	-0.0346	-0.3505
NEU-OPE	-0.17	-0.2868	-0.1419	-0.1750	0.0034
AGR-CON	0.43	0.4263	0.1008	0.6360	0.4169
AGR-OPE	0.21	-0.4248	-0.0785	-0.4895	-0.1777
CON-OPE	0.20	0.0866	0.2049	-0.6385	-0.6023

Table 8: Comparison of Correlation Values from van der Linden et al. (2010) and predictd personality traits. (Note. Red font indicates the difference in the direction of correlation, and Green font indicates the consistencies of direction)

5.4 Influence of Annotation for traits correlations

Based on the previous section, discrepancies are observed between the findings from van der Linden et al. (2010), Rushton and Irwing (2008), and the results from this study. These variations could potentially be attributed to the differential capabilities of the models in capturing the correlations among traits. However, there are also divergences in correlations derived from the datasets (i.e., PAN 2015, PANDORA) compared to those reported in the meta-analysis, as shown in Table 9 below.

For PAN 2015 dataset, the personality traits were self-assessed with the BFI-10 online test (Rangel et al., 2015), while for the PANDORA dataset, the process of obtaining Big 5 labels was complex. In particular, the PANDORA dataset also adopts the results from users' self-accessed tests; however, the test results are from 12 different questionaries, resulting in a wide array of reporting formats. Thus, the normalisation of these scores presented a significant challenge (i.e. the varying nomenclature for traits and the diverse



Traits	van der Linden et al. (2010)	PAN 2015		PANDORA	
		RoBERTa	Bi-LSTM	RoBERTa	Bi-LSTM
EXT-NEU	−0.36	0.3773	0.3951	-0.7611	-0.5870
EXT-AGR	0.26	0.3380	0.2348	-0.6480	-0.4755
EXT-CON	0.29	0.3345	0.1884	-0.1260	0.1177
EXT-OPE	0.43	-0.0567	-0.1398	0.4436	0.3267
NEU-AGR	-0.36	0.5632	0.3093	0.3321	-0.0515
NEU-CON	-0.43	-0.0729	-0.1796	-0.0346	-0.3505
NEU-OPE	-0.17	-0.2868	-0.1419	-0.1750	0.0034
AGR-CON	0.43	0.4263	0.1008	0.6360	0.4169
AGR-OPE	0.21	-0.4248	-0.0785	-0.4895	-0.1777
CON-OPE	0.20	0.0866	0.2049	-0.6385	-0.6023

Table 9: Comparison of Correlation Values from van der Linden et al. (2010), Rushton & Irwing (2008), PAN 2015, PANDORA and Sampled Pandora. (Note. Red font indicates the difference in the direction of correlation)

methods of score reporting, which included raw scores, percentages, or percentiles). Additionally, the scores could be either numeric or descriptive, each with its own set of ranges or descriptors specific to each test (Gjurković et al., 2020). The extraction process was semi-automatic, necessitating manual verification, which is a subjective process and can vary greatly among different annotators. Consequently, the correlations among traits in the original datasets may be influenced by these factors, leading to differences in the observed correlations compared to those reported in the meta-analysis studies.

As shown in Table 6 and Table 7, both the RoBERTa and Bi-LSTM models were able to capture parts of the relationships (i.e. Negative and positive correlations) presented in the original datasets. However, the correlations among traits in the original datasets differed from those reported in the meta-analysis studies. This discrepancy suggests that the models were primarily learning the correlations present in the annotated data, which may not necessarily align with the correlations found in the broader psychological literature.

Moreover, the diversity of data in the two datasets, influenced by the annotation process, could also have affected the performance of the models as well. As we discussed in the previous section, the data diversity may infect the models' capability to capture the variance in personality traits and RMSE of predictions, while the use of various annotation methods may directly influence the data diversity of trait values.

While machine learning models could learn and capture the correlations present in the data they are trained on, the influence of annotation on trait correlations cannot be overlooked. The annotation process, including the method of data collection and normalisation, can introduce variability and potential bias into the data, which can subsequently affect the correlations among traits and the performance of the models. Therefore, it is crucial to consider the influence of annotation when interpreting the results of studies on personality trait prediction.

5.5 Limitations

Despite the findings and discussions mentioned above, this study has several limitations. Firstly, there are multiple possible improvements for the data collection; for instance, this study relies on two public datasets which may not fully represent the personalities of populations. The correlations of personality traits in the datasets have divergences with findings from multiple studies (Rushton & Irwing, 2008; van der Linden et al., 2010), which may suggest the facts of being underrepresented. Furthermore, the limitations of inputs length for RoBERTa, the data restructuring for PAN 2015 dataset may influence the performance of prediction; originally, the tweets from one user are stored in one list, which may provide a more comprehensive information when predicting the personality, the separation of tweets may result in the loss of information. Moreover, for the data preprocessing part, it is challenging to transform the texts into "cleaned text" since there are multiple abbreviations and special words that Bi-LSTM could not understand, which may limit the performance of Bi-LSTM. Also, this study does not explore the performance of more complex deep learning model architectures like AttRCNN-CNNs due to limitations on computational resources. Regarding model training, due to time constraints and hardware limitations, the grid search did not encompass a broad range of potential values for hyperparameters.

5.6 Ethic Considerations

The datasets used in this study include publicly available text data, and it is important to follow the terms of use for both datasets. In this paper, we outline the structures of the datasets, which include a portion of the information. However, we anonymise the information that could potentially identify individual users, which is a crucial part when using or showing the datasets.

Another ethical consideration is the potential biases in the predictions made by machine learning models. These models are trained to predict users' personality traits based on UGTs. However, biases can be introduced in various ways, such as through data collection, the design or training of the models, or the interpretation of the predictions. For example, if the training data predominantly comes from a specific demographic group, the models may be more accurate in predicting the personality traits of individuals from that group and less accurate for others. This could lead to unfair outcomes, where specific individuals or groups are systematically disadvantaged by the predictions of the models.

In the future applications, such as employment, those biases may potentially lead to individuals being unfairly judged based on their predicted personality traits. For instance, if an individual is predicted to have low conscientiousness and this prediction is used to deny them a job opportunity, this could be considered unfair. To reduce the risks, it is vital to consider the potential biases when using automatic personality detection. One approach could be to use personality computation at an aggregate level rather than at the individual level. This could help reduce the risk of unfairly judging individuals based on their predicted personality traits.

6. Conclusion and Further Research

In conclusion, this study provides a comparative analysis of two machine learning models, RoBERTa and Bi-LSTM, in predicting the Big Five personality traits from text data and capturing correlations among personality traits. The results demonstrate the better performance of the RoBERTa model on both datasets, as evidenced by lower RMSE and higher R^2 values across all personality traits. Also, RoBERTa could better capture the correlations compared to the correlations in the original datasets since it provided more similar correlations with the original one.

Interestingly, the study also indicatess that both models, RoBERTa and Bi-LSTM, could capture most of the directions of correlations (i.e., negative or positive) among personality traits. However, RoBERTa exhibits a more consistent performance across the two datasets, with fewer divergences in the captured correlations. This suggests that RoBERTa may be more robust in learning and generalising the intercorrelations among personality traits.

Furthermore, the study highlights the influence of datasets on the performance of the models. The models perform differently on the PAN 2015 and PANDORA datasets, both in terms of predicting personality traits and capturing correlations. The higher RMSE observed on the PANDORA dataset may be attributed to its higher data variability, which presents a more challenging task for the models. Meanwhile, the correlations captured by the models are highly influenced by the correlations present in the original datasets.

Furthermore, all observed divergences in the direction of correlations within this study occurred in instances where the correlations were weak. This observation suggests that the models may have difficulties in accurately capturing the relationship between variables when the correlation is not strong.

For future research, it is possible to explore the performance of models with additional features or combined architecture, for instance, adding demographic features or using architecture like AttRCNN-CNNs. Furthermore, it is also essential to explore the performances of the machine learning methods on various datasets to ensure the capabilities of models in capturing correlations among personality traits. Lastly, it is also possible to explore the role of R^2 values in measuring the performance of models and improving the R^2 values of models since they are relatively low in this study.

A. Appendix

A.1 Github link

https://github.com/coxon1/Personality-dection_RoBERT_Bi-LSTM

Bibliography

Adi, G. Y. N., Tandio, M. H., Ong, V., & Suhartono, D. (2018). Optimisation for automatic personality recognition on twitter in bahasa indonesia. *Procedia Computer Science*, *135*, 473–480.

Ahmad, H., Asghar, M. U., Asghar, M. Z., Khan, A. A., & Mosavi, A. H. (2021). A hybrid deep learning technique for personality trait classification from text. *IEEE Access*, *9*, 146214–146232.

Allik, J., & McCrae, R. R. (2002). A five-factor theory perspective. *The five-factor model of personality across cultures*, 303–322.

Arijanto, J. E., Geraldy, S., Tania, C., & Suhartono, D. (2021). Personality prediction based on text analytics using bidirectional encoder representations from transformers from english twitter dataset. *International Journal of Fuzzy Logic and Intelligent Systems*, *21*(3), 310–318. https://doi.org/10.5391/ijfis.2021.21.3.310

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26.

Bird, S., Loper, E., & Klein, E. (2009). Natural language processing with python. *Proceedings of the 2009 Annual Conference on Python for High-Energy Physics and Astrophysics (CHEP '09)*, 11–18.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei, D., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Costa Jr, P. T., & McCrae, R. R. (2008). *The revised neo personality inventory (neo-pi-r)*. Sage Publications, Inc.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 4171–4186.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology*, *93*(5), 880–896.

Digman, J. M. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology*, *73*(6), 1246–1256.

Esterwood, C., & Robert, L. P. (2020). Personality in healthcare human robot interaction (h-hri) a literature review and brief critique. *Proceedings of the 8th International Conference on Human-Agent Interaction*, 87–95.

Fang, Q., Giachanou, A., Bagheri, A., Boeschoten, L., van Kesteren, E.-J., Kamalabad, M. S., & Oberski, D. L. (2022). On text-based personality computing: Challenges and future directions. *arXiv preprint arXiv:2212.06711*.

Gjurković, M., Karan, M., Vukojević, I., Bošnjak, M., & Šnajder, J. (2020). Pandora talks: Personality and demographics on reddit. *arXiv preprint arXiv:2004.04460*.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology*, *2*, 141–165.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, *37*(6), 504–528.

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. In *Handbook of personality psychology* (pp. 795–824). Elsevier.

Guntuku, S. C., Preoţiuc-Pietro, D., Eichstaedt, J. C., & Ungar, L. H. (2020). What twitter profile and posted images reveal about depression and anxiety. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 236–246.

Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia computer science*, *17*, 26–32.

Hall, M., & Caton, S. (2017). Am i who i say i am? unobtrusive self-representation and personality recognition on facebook. *PloS one*, *12*(9), e0184417.

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 187–196.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*(2), 161–172.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big five inventory. *Journal of personality and social psychology*, *61*(3), 524.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of personality psychology* (pp. 114–158). Elsevier.

John, O. P., Robins, R. W., & Pervin, L. A. (2010). Handbook of personality: Theory and research. Guilford Press.

John, O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, *2*(1999), 102–138.

Kreuter, A., Sassenberg, K., & Klinger, R. (2022). Items from psychometric tests as training data for personality profiling models of twitter users. *arXiv preprint arXiv:2202.10415*.

Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist*, *64*(4), 241–256.

Li, M., Liu, H., Wu, B., & Bai, T. (2022). Language style matters: Personality prediction from textual styles learning. *2022 IEEE International Conference on Knowledge Graph (ICKG)*, 141–148.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *Advances in neural information processing systems*, *33*, 1877–1901.

Lucas, R. E., & Baird, B. M. (2004). Extraversion and emotional reactivity. *Journal of Personality and Social Psychology*, *86*(3), 473–485.

Ma, W., Cui, Y., Si, C., Liu, T., Wang, S., & Hu, G. (2020). Charbert: Character-aware pre-trained language model. *arXiv preprint arXiv:2011.01513*.

Matthews, G., Deary, I. J., & Whiteman, M. C. (2003). *Personality traits*. Cambridge University Press.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90.

McCrae, R. R., & Sutin, A. R. (2009). Openness to experience. *Handbook of Individual Differences in Social Behavior*, 257–273.

Miller, D. (2019). Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Myers, I. B. (1962). *The myers-briggs type indicator: Manual*. Consulting Psychologists Press.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Paulhus, D. L., & Vazire, S. (2007). The self-report method. *Handbook of research methods in personality psychology*, *1*, 224–239.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001).

Phan, L. V., & Rauthmann, J. F. (2021). Personality computing: New frontiers in personality assessment. *Social and Personality Psychology Compass*, *15*(7), e12624.

Rangel, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd author profiling task at pan 2015. *CLEF 2015 Labs and Workshops, Notebook Papers*, 1–8.

Ren, Z., Shen, Q., Diao, X., & Xu, H. (2021). A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, *58*(3), 102532.

Roberta-base. (n.d.). Roberta-base · hugging face [[online] Available: https://huggingface.co/roberta-base (Accessed: 20 June 2023)].

Rushton, J. P., & Irwing, P. (2008). A general factor of personality (gfp) from two meta-analyses of the big five: And. *Personality and Individual Differences*, *45*(7), 679–683.

Samuel, D. B., & Widiger, T. A. (2008). A meta-analytic review of the relationships between the five-factor model and dsm-iv-tr personality disorders: A facet level analysis. *Clinical Psychology Review*, *28*(8), 1326–1342.

Sanchez-Roige, S., Gray, J. C., MacKillop, J., Chen, C.-H., & Palmer, A. A. (2018). The genetics of human personality. *Genes, Brain and Behavior*, *17*(3), e12439.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

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., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791.

Shen, T., Jia, J., Li, Y., Ma, Y., Bu, Y., Wang, H., Zhu, L., & Hall, W. (2020). Peia: Personality and emotion integrated attentive model for music recommendation on social media platforms. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 206–213.

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Ohana, S., Lehmann, J., Clark, M., et al. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680–17687.

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Pentland, A., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*, *34*(5), 613–631.

Štajner, S., & Yenikent, S. S. (2020). A survey of automatic personality detection from texts. *Proceedings of the 28th International Conference on Computational Linguistics*, 6284–6295.

Sun, X., Liu, B., Cao, J., Luo, J., & Shen, X. (2018). Who am i? personality detection based on deep learning for texts. *2018 IEEE international conference on communications (ICC)*, 1–6.

Tandera, T., Suhartono, D., Wongso, R., & Prasetio, Y. L. (2017). Personality prediction system from facebook users. *Procedia computer science*, *116*, 604–611.

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327.

Xue, D., Wu, L., Hong, Z.-R., Guo, S., Gao, L., Wu, Z., Shen, J., & Sun, J. (2018). Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, *48*, 4232–4246.

Yang, H.-C., & Huang, Z.-R. (2019). Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems*, *165*, 157–168.

Yu, J., & Markov, I. (2017). Deep learning based personality recognition from facebook status updates. *2017 IEEE 8th international conference on awareness science and technology (iCAST)*, 383–387.