TEXT BASED PERSONALITY RECOGNITION BASED ON USER'S CONTENT

STUDENT NAME: Mujahid Ahmad
ENROLLMENT NO: 01-249212-009
SUPERVISOR: Dr. Muhammad Asfand E Yar

A thesis submitted in fulfillment of the requirements for the award of a degree of Masters of Science (Computer Science)

Department of Computer Science

BAHRIA UNIVERSITY ISLAMABAD

OCTOBER 2023

Approval of Examination

Scholar Name: Mujahid Ahmad
Registration Number: 41978
Enrollment: 01-249212-009
Program of Study: MS (Data Science)
Thesis Title: Text Based Personality Recognition Based on User's Content

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index 7%. that is within the permissible limit set by the HEC for the MS/M.Phil degree thesis. I have also found the thesis in a format recognized by the BU for the MS/M.Phil thesis.

Principal Supervisor Name: Dr. Muhammad Asfand e yar

Principal Supervisor Signature: _____

Date: 24-Oct-2023

Author's Declaration

I, Mujahid Ahmad hereby state that my MS thesis is my own work and has not been submitted previously by me for taking any degree from Bahria University or anywhere else in the country/world. At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS/M.Phil degree.

Mujahid Ahmad

24-Oct-2023

Plagiarism Undertaking

I, solemnly declare that the research work presented in the thesis titled Text Based Personality Recognition Based on User's Content is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me. I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, as an Author of the above-titled thesis, I declare that no portion of my view has been plagiarized and any material used as reference is properly referred to / cited.

I undertake that if I am found guilty of any formal plagiarism in the above-titled thesis even after the award of the MS degree, the university reserves the right to withdraw/revoke my MS degree and that HEC and the The university has the right to publish my name on the HEC / University website on which terms of scholars are placed who submitted plagiarized thesis.

Mujahid Ahmad

24-Oct-2023

Dedication

This thesis is dedicated to my brother, Munir Ahmad.

To Munir,

Your unwavering support, boundless encouragement, and relentless belief in me have been the driving force behind the completion of this thesis. Through the challenges and uncertainties, you stood by my side, never letting me give up, and continually pushing me to reach my fullest potential.

Your faith in my abilities has been a guiding light, and your presence in my life is a constant source of inspiration. This achievement is as much yours as it is mine, for your support has been instrumental in making it a reality.

With heartfelt gratitude and love,

Mujahid Ahmad

October 24, 2023

Acknowledgements

I would like to express my sincere gratitude and appreciation to the following individuals and organizations for their invaluable support, guidance, and contributions throughout the Text-Based Personality Recognition Based on User's content. I am deeply thankful to my advisor, Dr. Muhammad Asfand Yar, for their unwavering support, expert guidance, and insightful feedback. Their mentorship has been instrumental in the successful completion of this thesis. I extend my heartfelt thanks to my family for their constant encouragement and belief in my abilities. My friends have also been a source of motivation and support during this journey. I am grateful to my colleagues/classmates for their collaborative efforts, meaningful discussions, and exchanging ideas that enriched this work. I acknowledge Bahria University for providing access to resources, facilities, and a conducive environment for research and learning. This project would not have been possible without the collective support of these individuals and institutions. Thank you for your unwavering encouragement and assistance.

Mujahid Ahmad

24-Oct-2023

LIST OF SYMBOLS

EXT	–	Extraversion
NEU	–	Neuroticism
AGR	–	Agreeableness
CON	–	Conscientiousness
OPN	–	Openness
RNN	–	Recurrent Neural Network
BiLSTM	–	Bidirectional Long Short-Term Memory
BiGRU	–	Bidirectional Gated Recurrent Unit
I/E	–	Extraversion (E) or Introversion (I)
S/I	–	Sensing (S) or INtuition (N)
T/F	–	Thinking (T) or Feeling (F)
J/P	–	Judging (J) or Perceiving (P)

Abstract

The potential applications of personality prediction from textual data in psychology, marketing, and human-computer interaction have sparked considerable attention in recent years. While previous research has provided useful insights, this work takes a novel approach to personality prediction by combining the power of powerful transformer-based models such as BigBird, Albert, and DistilBERT with NLP statistical characteristics. Notably, these cutting-edge models have never been used in this context before. The goal of this study is to thoroughly examine and compare the performance of these advanced models, enhanced with NLP statistical features, vs. conventional methods in predicting personality traits across varied textual datasets such as the Facebook dataset and the essay dataset. By doing so, the study hopes to shed light on the untapped potential and challenges inherent in using transformer-based models and NLP statistics for personality trait prediction, advancing our understanding of their capabilities and the advantages they offer over established techniques. In this study, we used two classifiers, BiGRU and BiLSTM, to classify five personality traits of Big 5 personality trait model using Facebook and essay datasets. When combined with NLP statistical features and BiLSTM, BigBird achieves F1-scores of 0.82, 0.76, 0.74, 0.84, and 0.81 for the traits EXT, NEU, AGR, CON, and OPN, respectively, with accuracies of 85.16%, 87.39%, 92.35%, 98.48%, and 98.33% on the Facebook dataset. These findings illustrate the power of advanced transformer-based models augmented with NLP statistics in predicting personality across diverse datasets. Our evaluation also includes accuracy and F1-score results for each attribute and dataset, allowing us to provide a full assessment of our models' performance. This study adds to the growing field of personality prediction by bringing advance approaches and emphasizing the efficiency of sophisticated transformer-based models in comprehending human behavior through textual data.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURE

CHAPTER 1

INTRODUCTION

1.1 Introduction

The importance of comprehending the complexities of human personality increases in a society that is rapidly changing and becoming more connected thanks to the digital age. Our personalities, which are made up of an intricate tapestry of feelings, thoughts, and actions, are not just the result of nature and nurture but also a defining characteristic that affects how we interact with the outside world. Deciphering personality features plays a crucial role in allowing successful communication and establishing deeper connections among people in an age where the internet and social media smoothly cross geographical boundaries. It is impossible to emphasize the enormous influence that personality has on many aspects of life. Our personalities serve as the lighthouse guiding us through the intricacies of life, from influencing life choices to determining degrees of satisfaction and involvement. However identifying these innate characteristics has typically proved difficult, requiring in-depth psychological evaluations and observations. However, the current digital environment has made text-based communication a fascinating new way to explore personality. The complex internet and the rise of social media have brought people from different backgrounds together like never before in today's globalized society. Our online persona is constructed through the words we write, the emotions we convey, and the subjects we discuss. As a result, there is an unmistakable connection between a person's character and their online behavior. In addition to piquing academics' interest, this realization has sparked the creation of cutting-edge approaches for extrapolating personality traits from the large body of internet debate.

Explore the world of personality recognition, a field that combines artificial intelligence, natural language processing, and psychology. These disciplines have come together to create fresh methods for solving the complex textual personality challenge. Researchers have started to reveal the layers

of meaning buried inside our words by leveraging the power of pre-trained models, machine learning algorithms, and deep learning architectures. The mutually beneficial interaction of these technological wonders with the subtleties of human language has made it possible to comprehend personality traits at a deeper level. However, the effects of personality recognition go far beyond personal interest. Organizations can benefit greatly from such efforts by learning priceless lessons. The applications are numerous, ranging from customizing marketing techniques to aligning with distinctive client personas to screening individuals throughout the employment process. A better understanding of customer temperament can help develop meaningful conversations, eventually leading to improved business partnerships. We set out on a tour across the field of text-based personality recognition in the pages that follow. We will examine one of these frameworks where the Big Five converge with AI and machine learning by delving into the domains of well-established personality models like the Myers-Briggs Type Indicator (MBTI), the Big Five, and others. We will examine the mechanisms that enable us to extract personality traits from written text using pre-trained models and sophisticated algorithms.

The confluence of human psychology and technology prowess will be on display as we travel this thrilling landscape, revealing the intricate web of personalities as each keystroke is scrutinized. By the time we're done, I'm hoping that the complex dance between human expression and computer interpretation will be more obvious and that the potential for improving comprehension and communication will be more tangible than ever.

1.2 Personality Recognition Models

The four models used to determine personality type are the MBTI, The Big Five, 16 Personalities, and Type A, B, C, and D Personalities.

1.2.1 MBIT Model

Based on Carl Jung's theory, the MBTI was created in the 1920s to identify personality types according to where you fall in four categories:

- I/E
- S/I
- T/F
- J/P

A set of questions called the MBTI tests is used to gauge psychological preferences in how individuals view the world and make decisions:

- Favorite world: Do you prefer to concentrate on the outside or the inside of the world? This is referred to as I/E.
- Information: Do you prefer to interpret and add meaning or do you focus more on the foundational knowledge you take in? This is defined as I/N.
- Decisions: Do you favor considering people and unique circumstances before consistency and logic while making choices? This term is called T/F.
- Structure: Do you like to make decisions upfront when interacting with the external world, or do you choose to be flexible and open to unique information and options? This is defined as J/P.

Figure 1.1: MBTI Personality Trait

1.2.2 16 Personalities Model

Like the MBTI, 16 Personalities is a framework for identifying different personality types. Contrary to Myers-Briggs, anyone can take the 16 Personalities test for free online at the 16 Personalities website.
In this model, different Myers-Briggs-like personality types are divided into four classes.

- The Analysts: INTJ, INTP, ENTJ, ENTP which is called Architect, Logician, Commander, and Debater respectively.
- The Diplomats: INFJ, INFP, ENFJ, ENFP, which is defined as Advocate, Mediator, Protagonist, and Campaigner respectively.

- The Sentinels: ISTJ, ISFJ, ESTJ, ESFJ which is defined as Logistician, Defender, Executive, and Consul respectively.
- The Explorers: ISTP, ISFP, ESTP, ESFP called Virtuoso, Adventurer, Entrepreneur, and Entertainer respectively.

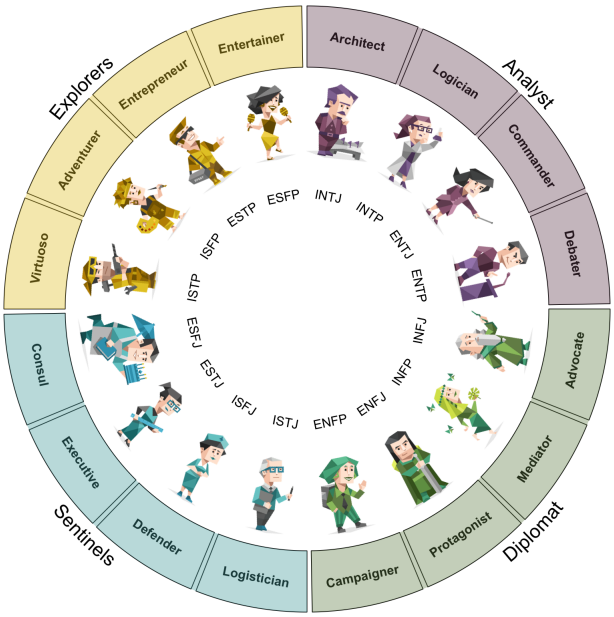

Figure 1.2: 16 Personality Model

1.2.3 The Big Five Model

The acronym CANOE is used to refer to the big 5 personality types, that were created in the 1980s.:

- CON (conscientiousness)
- AGR (agreeableness)
- NEU (neuroticism)
- OPE (Openness to experience)
- EXT (Extraversion)

Studies have shown the Big 5 Five test to be a reliable predictor, frequently used in intellectual psychological personality research.

Figure 1.3: The Big Five Personality Model

1.2.4 Type A, B, C and D Personalities Model

In 1976, Meyer Friedman and Ray Rosenman first distinguished between the Type A and Type B personalities, Type A is sensitive to stress, and Type B is more laid back. Over time, their framework has been expanded to include Type C and Type D personalities, which are all described as follows:

- Type A: They are natural leaders who try to exert the most control. They are called Directors, Overachievers, or Go-Getters.

- Type B: The Socializer or The Peacemaker, is the antithesis of Type A personalities. In good circumstances, these gregarious people are fun to be around, but they occasionally border on being needy.

- Type C: When they feel out of control, they can easily become overwhelmed because they use logical reasoning to make sense of the world. They are called The analyst or the thinker.

- Type D: They are aware of their feelings and may find it difficult to be upbeat, which makes them more like Type B personalities. They are mysterious and sensitive, experiencing happiness and joy more intensely than other people while also being more prone to anxiety and depression. They are called supporters or philosophers.

1.3 Background Motivation

Personality has long been a source of interest and research in the complex field of human psychology. The distinctive pattern of each person's identity is defined by their personality, which is an intricate fusion of their thought processes, emotional reactions, and behavioral tendencies. As scientists and psychologists have worked to comprehend, classify, and forecast these complex features across time, a variety of personality models have come to be. The Big Five Personality Model, which includes five key dimensions that collectively reflect the core of a person's disposition, stands out among them as a cornerstone. The opportunities for researching and utilizing personality traits have greatly increased as the digital era develops. The quick development of Natural Language Processing (NLP) and the overlap between personality detection and this subject have sparked this expansion. The goal of NLP, a branch of artificial intelligence, is to give machines the capacity to comprehend, decipher, and produce human language. This potential has created hitherto unexplored opportunities for revealing the subtleties of personality buried within language, especially when combined with the vast textual data readily available online. The nexus of personality recognition and NLP offers a fresh way to interpret the psychological characteristics of a person from their textual expressions. This field combines the accuracy of machine learning algorithms with the intricacies of linguistic patterns. Researchers and practitioners can now reveal insights about people's personality qualities that were previously hidden beneath layers of words by utilizing various computational methodologies. In a variety of fields, personality detection from text is quite important. This strategy goes beyond the restrictions of conventional assessment techniques in psychology, which frequently employ biased self-report surveys. It may be possible for researchers to develop a more accurate grasp of people's personalities by studying the spontaneous text. Ingenious therapeutic applications have also been made possible by the blending of psychology and NLP, allowing mental health experts to monitor changes in patients' personality features through textual expressions. The effects of personality recognition go well beyond the field of psychology. For instance, in marketing, being able to infer consumer personalities from their online interactions enables the creation of highly targeted advertising campaigns. Marketers may increase customer engagement and conversion rates by personalizing messaging to fit the personality qualities of their target audience. Personality recognition aids in both human-computer interaction and communication. Virtual assistants, chatbots, and recommendation systems can all benefit from systems that can adjust to the personalities of their users to offer more effective and individualized experiences. The com-

plex procedure of feature extraction and categorization utilizing sophisticated models forms the core of your research. You may bridge the gap between the complexity of human expression and the accuracy of AI analysis by utilizing the power of pre-trained transformers and incorporating statistical information. The bridge is provided by the attention-based models you use, which make it possible to spot tiny patterns and connections within the text that point to underlying personality traits.

1.4 Problem Statement

In the age of online personalization and abundant user-generated text data, there is a need to create an automated system that uses advanced AI models and NLP techniques to accurately discern an individual's personality traits from textual content such as blog posts, essays, and social media posts. This study aims to develop a personality recognition system capable of extracting personality insights from textual user material.

1.5 Research Objective

- Investigate the applicability of advanced transformer-based models such as BigBird, Albert, and DistilBERT when used as feature extractors for predicting personality traits.
- Examine how incorporating NLP statistical information alongside transformer-based models affects the accuracy and F1-score of personality trait prediction.
- To design a system that improves the accuracy and F1 score for each five traits in the Big Five model

1.6 Significance of Research

The Big Five Personality Model was the subject of text-based personality recognition research because it has the potential to improve communication dynamics, enable more individualized interactions, and advance our understanding of human behavior. This study's significance can be seen in numerous striking ways

1.6.1 Increasing Psychological Understanding

By identifying personality qualities in text, this study advances psychological knowledge of the intricate relationship between linguistic expression and

psychological characteristics. It fills the gap between conventional self-report tests and impulsive online conversation, perhaps improving our comprehension of how people express their personalities in varied circumstances.

1.6.2 Improving Human-Computer Interaction

Accurate text-based personality detection could completely transform how people and computers communicate. Systems that can recognize and react to users' personality features, such as chatbots and virtual assistants, can provide more individualized and interesting experiences. This results in increased user engagement and pleasure.

1.6.3 Marketing Strategies Can Change

Enhanced personality recognition accuracy can alter marketing tactics. The personalities of consumers can be deduced from their online interactions, allowing marketers to create messages and adverts that appeal to specific tastes. This tailored strategy may result in greater conversion rates and more fruitful interaction.

1.6.4 Enriching Mental Health Assessment

In the field of mental health, precise personality recognition from text can offer insightful information about people's emotional states and psychological health. This may make it easier for mental health specialists to monitor changes, spot possible problems, and provide more focused interventions.

1.6.5 Personalized recommendation systems

A greater comprehension of user personalities can help recommendation systems in a variety of fields, including entertainment, material, and products. Increasing the precision of personality detection can produce recommendations that are more in line with personal preferences.

1.6.6 Making Well-Informed Judgements in Organisations

Businesses can use text-based personality recognition insights to help them make better judgments. Understanding personality features can help with better matches and interactions, which can lead to better results in anything from hiring employees to customer engagement.

CHAPTER 2

RELATED WORK

2.0.1 Feature Extraction Technique

David Stillwell's 2007 myPersonality app pioneered psychometric testing on Facebook, using digital platforms to generate an extensive database of research and insight into personality traits [1]. Research into linguistic patterns and their connection to personality traits underscores the potential of language analysis to uncover psychological differences and emphasizes the complex interplay between language expression and individual specificity [2]. The study presents the TF-IGM technique as an alternative to TF-IDF for text classification and shows higher accuracy by integrating term frequency and inverse group frequency [3]. Research on sentiment analysis via Twitter highlights the challenges of concise and real-time news formats and presents strategies that include lexical-based and machine-learning techniques [4]. Innovations in Transformer models such as Big Bird, ALBERT, and DistilBERT address memory and computational efficiency issues while maintaining or improving word problem performance [5][6][7]. Text mining and sentiment analysis are gaining importance due to the abundance of data from social networks; various approaches are being explored to address the challenges of noisy and unstructured text data [20]. By examining emotional expression in social media, researchers link personality traits to emotional disclosure in status updates and reveal associations between individual traits and emotional expression [8]. Tools such as the NRC Emotion Lexicon enhance sentiment analysis by associating words with emotional tendencies, allowing for sophisticated computational analysis of emotional expression in different textual contexts [9]. The introduction of LIWC2007 enriches text analysis by improving linguistic and psychological content analysis skills, thus contributing to the understanding of psychological concepts in text data [10]. Flesch's seminal work on text readability [11] addressed the urgent need for a quantitative measure to objectively assess the complexity of written texts. His formula calculated a readability score based on sentence length and the number of syllables per word.

This assessment provided a numerical index that quantified understandability among different audiences. By developing a systematic framework for assessing text complexity, Flesch laid the groundwork for further research in the field of readability analysis. The Flesch Reading Ease Score, a result of his work, is still a widely used metric for assessing accessibility. and the suitability of the content in various areas, from educational materials to legal documents. Flesch's contribution had a profound impact on linguistics, psychology, and education, and provided a quantitative understanding of text readability that remains relevant and influential today.

2.0.2 Machine Learning Approach

In recent years, the study of personality traits through machine learning approaches has gained importance. Various methods and techniques have been used to analyze personality traits from textual data.

A machine learning approach was used for personality trait analysis, which includes Named Entity Identification (NEI) and feature engineering methods such as TF-IDF and word embeddings such as Word2vec. The study leveraged the Myers-Briggs type indicator dataset and used techniques ranging from data pre-processing and feature engineering to classification and ensemble learning methods. Algorithms such as XGBoost, bagging (random forest), and stacking were used in ensemble learning, and the best model was based on comparisons of different ensemble results [12]. Efforts have been made to improve the efficiency of the model using XGBoost and study the impact of the HEXACO model on business success. The methodology involved data-level resampling, k-fold cross-validation, and various machine learning classifiers to determine the most effective ones [13]. Supervised machine learning algorithms were used to predict users' personality traits based on the entered text. Extensive preprocessing and feature engineering methods were used, including stemming, URL and hashtag stripping, binarization, stopword stripping, and TF-IDF feature selection. Several classifiers including RF (Random Forest), XGBoost, Gradient Descent, LR, ANN model, and SVM were compared to assess model accuracy [14].XGBoost was used to categorize personality traits from user text. The methodology included data collection, resampling, preprocessing, feature selection, and classification using the MBTI model. A performance comparison was performed between XGBoost and other classifiers, supported by various evaluation metrics [15].

2.0.3 Deep Learning Approach

Research into personality recognition through deep learning techniques has led to various methodologies and approaches and contributed to the advancement of the field.
A personality recognition model was developed using deep learning techniques using convolutional neural networks (CNN) and incorporating new features extracted using Linguistic Query Analysis and Word Count (LIWC). The "MyPersonality" data set was used and LIWC2015 was used for feature extraction. The CNN model, configured with the same parameters as stochastic gradient descent (SG), formed a binary classification model to achieve personality detection [16]. A personality profile was created using Computational Psychology by assessment on the MBTI scale. Pre-processing steps included removing hyperlinks, numbers, and punctuation while employing derivation techniques such as WordNet Lemmatizer and Lancaster Stemmer. A feature vector was created by combining features from TF-IDF, EmoSenticNet (10 emotions), LIWC, and ConceptNet (300 floating point numbers). The classifiers SVM, Neural Networks, and Naive Bayes were trained and evaluated using the MBTI dataset with a splitting ratio of 70:30. The results showed that SVM achieved the best accuracy of 86.27% [17].SEPRNN uses deep learning and contextual learning for efficient recognition of multi-tagged personality traits from text data and shows advances in semantics-based personality recognition [10]. Researchers integrate emojis into personality recognition models using bidirectional long-term and short-term memory (BiLSTM) and attentional mechanisms to improve personality recognition based on text and emoji information [18]. Deep learning algorithms including fully connected neural network (FC), convolutional neural network (CNN), and recurring neural network (RNN) were used for personality detection. Word embedding and network architecture were key components. A Skip-Gram-based word embedding matrix was pre-trained. Individual networks were trained for each of the five personality qualities. The optimal results were obtained with a convolution architecture using average pooling and achieved a precision of 60.06.5%. CNN, RNN, and FC were identified as effective feature extractors, while recurrent architecture, trigram, and bigram gave no better results [19]. A hybrid deep learning model that aims to classify text based on specific personality traits. Tokenization, stop word stripping, and lowercase were applied to the text. The model combined the DNN-CNN+LSTM architecture and used an embedding layer for word representation, CNN for feature extraction, LSTM for long-term information learning, and a SoftMax layer for classification [20]. A unique approach involving CNN and the AdaBoost method was explored to

explore the potential of combining inputs from filters of different lengths for personality identification. Datasets from YouTube personalities and stream-of-consciousness studies were used. Word embedding based on the Skip-Gram model was used to extract local features. AdaBoost was used to scale the classifier with different n-grams, highlighting the role of pooling and dropping strategies [21].

2.0.4 Pre-trained Transformers

A multi-model deep learning architecture integrates NLP functions with pre-trained transformers such as BERT, RoBERTa, and XLNet. Preprocessing includes removing URLs, symbols, and emoticons, followed by English translation, lowercase, contraction expansion, stop word removal, and derivation. Feature extraction uses techniques such as word piece tokenization, token embedding, segment embedding, and position embedding, using CLS and SEP tokens to enhance the contextual meaning of the. This approach produces the best results for all major personality traits, including openness (70,85). %), conscientiousness (88.85%).49%), extraversion (81.17%), agreeableness (69.33%), and neuroticism (75.08%) [22]. The use of state-of-the-art DL-based NLP models meets the challenge of identifying and categorizing personality types using different fonts and text styles. Two datasets are indexed: MyPersonality (Facebook) and Essays (Penne and King). The research article proposes data-level and classifier-level fusion strategies to improve personality prediction performance. Pre-trained language models (Elmo, ULMFiT, BERT) are adopted, and combining the Essays and MyPersonality datasets further improves the proposed model [23]. It is not novel to predict personality traits using data from Facebook and Twitter. For example, [11] used an open-source Facebook personality dataset called MyPersonality, which contains 250 users' status data and attributes and maps to the huge five-personality model. The main feature extraction method is Linguistic Inquiry and Word Count (LIWC), which is a linguistic analytical tool that aids in the analysis of quantitative texts and provides a calculation number of words that have the meaning of categories based on a psychological dictionary.

2.1 Research Gap

There is gap in the usage of advanced deep learning models for feature extraction and classification in the context of personality evaluation in the current landscape of research, particularly when compared to the widespread use of machine learning models. Furthermore, while the Myers-Briggs Type

Indicator (MBTI) model is popular in some contexts, it is not widely accepted in the psychological community as a trustworthy instrument for personality assessment. This disparity highlights a serious research gap. The Big Five personality model, on the other hand, is widely recognized and approved among psychologists as a more advanced and respected framework for personality assessment. This disparity indicates a significant area of future research opportunity to investigate the potential of advanced deep learning models in personality assessment, with a focus on the validation and integration of the Big Five model within this framework, ultimately bridging the gap between cutting-edge technology and established psychological practices.

Table 2.1: Analytical Review Table 1. A

Ref	Year	Dataset	Problem Statement	Methodology	Technique	Results	Limitation
[12]	2020	MBTI	Analyze personality trait based on NEI	Stop words, data deduplication, NER, tokenization, lower casting, TF-IDF, word2vec	XGBoost, Bagging, Staking	S-I (95.79%) I-E (88.02%) T-F(77.69%) J-P(71.96%)	Limited machine learning classifier used.
[13]	2022	MBTI	The impact of HEXACO personality traits on business success.	re-sampling, k-fold cross-validation	XGBoost, ML classifiers	I/E, S/N (99% precision, accuracy) T/F, J/P (95% accuracy)	KNN classifier performed worse overall. The goal of personality recognition is to understand the relationship between a personality attribute and organizational achievement, not just for individuals.
[15]	2022	MBTI	Categories user text based on personality using XGBoost.	Extrapolating data into feature selection, data collection, resampling, and pre-processing.	XGBoost	I/E and S/N (99% accuracy)	The MBTI personality test was the subject of the work which is one of the four personality tests and one dataset is used to forecast personality attributes.
[14]	2022	MBTI	Review ML techniques for predicting users' personality attributes from the text. Feature engineering and pre-processing methods enhance performance against uneven personality traits in data.	Lemmatization, URLs, Hashtag, removal, binarizing, Stopwords removal, TF-IDF	Random Forest, XGboost, Gradient Descent, LR, KNN, SVM	SVM gives the best result for all MBTI traits. KNN (48.88% for J/P).	Researchers employed supervised machine learning classifiers to predict the personality qualities for the MBTI personality test.
[21]	2022	Essay by penne and king Youtube personality	Classifiers with the appropriate filter sizes, CNN is exploring the idea of utilizing the contributions of various filter size and gauging their ability to assess personality.	Skip-Gram Maximum pooling Dropout	AdaBoost	Essay Dataset Ext 61.25%, Neu 61.93%, Agre 59.02%, Open 60.16%, Con 64.63% Youtube Dataset Ext 62.11%, Neu 62.43%, Agr 60.23%, OpenEx 61.08%, Con 65.19%	The epochs are limited to 60, and while the accuracy has increased for the YouTube dataset, it has decreased for the Easy dataset.
[16]	2018	My-Personality	The personality detection model was built using deep learning for the Facebook dataset.	LIWC2015 to extract features	CNN	CNN OPEN(0.76%) MNB model EXT(0.58%) NEU(0.62%) AGR(0.59%)	Research is done only on the CNN algorithm there is more deep learning dataset other than the Facebook dataset on which personality recognition should be checked.

Table 2.2: Analytical Review Table 1. B

Ref	Year	Dataset	Problem Statement	Methodology	Technique	Results	Limitation
[17]	2018	MBTI	Create a profile of the individual by utilizing computational psychology to score on the MBTI scale.	Removal of Hyperlinks, numbers and punctuation from tweets, WordNet Lemmatizer, Lancaster Stemmer, Tweet Tokenizer, TF-IDF, EmoSenticNet(10 emotion), LIWC and ConceptNet	SVM Neural Network Navie bayes	S/N(90.45%) NN S/N(86.72%) NB S/N(88.27%) SVM	There are new algorithms in deep learning and pre-train model which we can use for model building.
[19]	2017	Subset of database computational personality recognition	On the task from the "Workshop on Computational Personality Recognition" deep learning algorithms such as FC, CNN, and RNN were evaluated.	word embedding and network architecture. Pre-trained word embedding matrix using the skip-gram method. 5 separate networks are trained to get each 5 personality traits	FC, CNN, RNN	CNN with average pooling is better than both the RNN and FC. Convolutional architecture with average pooling achieved the best results 60.06.5%.	The dataset used in this research paper experiment is related to Facebook status updates only
[20]	2021	MBTI	Classify text reviews into personality trait using hybrid deep learning model.	Data Lower casting, Eliminating stop words, Tokenization	DNN CNN+LSTM	Accuracy (88% for I-E, 91% for N-S, 85% for T-F,80% for J-P) Precision (88% for I-E, 91% for N-S, 85% T-F, 80% for J-P) F1-score (88% for I-E, 91% for N-S, 85% for T-F, and 80% for J-P)	Limited to English language and MBTI personality type.
[22]	2021	MyPersonality Twitter	Increase the amount of data for better classification and combine data from various sources. Review the model's performance and make a comparison to earlier research.	URLs, symbols, emoticons removed, English translation, Remove stop words, expand contractions,	BERT, RoBERTa, XLNet	OPEN (70.85%), CON (88.49%), EXT (81.17%), AGR (69.33%), NEU (75.08%), AVG (77.34%).	The model is limited to Facebook and Twitter data only.
[23]	2021	Eassys by penne and king Mypersonality	Regardless of the sources, textual styles, or psycholinguistic features used, DL-based models in NLP can be used to identify and categorize personality traits from the text.	Data level fusion Classifier level fusion	ELMo ULMFiT BERT	73.91% for MyPersonality 61.85% for essay dataset	Limited to three pre-trained transformer there are other pre-transformer on which is not used.

CHAPTER 3

METHODOLOGY

Our approach involves using two distinct datasets: the "essay" dataset [23] and the "Facebook" dataset [8]. Each dataset goes through a custom pre-processing sequence to ensure data quality and consistency. Both datasets go through a series of pre-processing steps to make the textual content accessible for further analysis. Fig 3.1 represents our proposed methodology to reach the desired output.

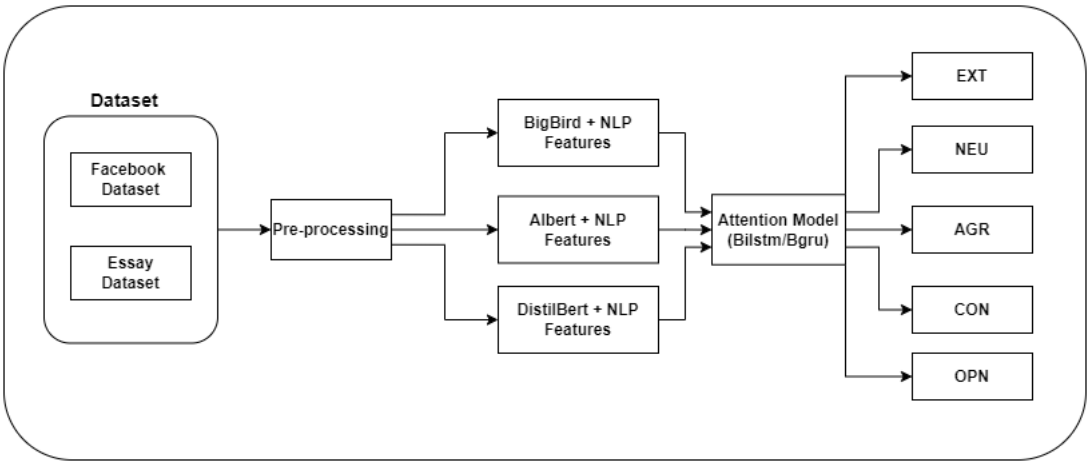

Figure 3.1: Proposed Methodology

The pre-processing phase serves to cleanse, normalize, and standardize the text data, thus creating a uniform basis for further exploration. From the pre-processed datasets, we extract key statistical characteristics of natural language processing (NLP).
In addition, we use the capabilities of advanced language models, namely BigBird, ALBERT, and DistilBERT, to extract complex semantic features from text [5] [6] [7]. This dual feature extraction process encapsulates the underlying

linguistic complexities in the data and provides a comprehensive representation for further analysis.
Feature-enriched data from each dataset are fed into the model for classification tasks. This phase includes the implementation of the bidirectional long-lived short-term memory (BiLSTM) models which are also used in the study [18] and the bidirectional gated recursive unit (BiGRU). These models have been carefully designed to capture complex dependencies and relationships within textual data. By taking advantage of the bi-directional nature of these models, we facilitate the recognition of contexts and patterns that are critical to the accurate classification of personality traits.

3.1 Dataset

Two different datasets were used to perform the analysis presented in this article, each providing unique information about the relationship between textual data and personality traits.

3.1.1 The myPersonality dataset

The first dataset, dubbed the "myPersonality dataset", includes a total of 250 Facebook users. This dataset comes from the myPersonality Project [8], a comprehensive initiative by Stillwell and Kosinski (2015) to study the associations between fingerprints and psychological traits. The myPersonality dataset is aligned with the framework of the Big Five personality traits paradigm and represents more than 9917 individual states associated with specific personality categories. As a subset of the larger dataset collected by the Facebook application, this subset provides a rich source of textual data for personality-oriented research purposes.
Figure 3.2 and Table 3.1 show visual and numerical representations of the distribution of the Big Five personality traits within the myPersonality dataset. These representations provide a thorough analysis of how many records or individuals fall into each personality trait's 'y' and 'n' categories. This data gives critical insights into the prevalence and diversity of these personality qualities among the 250 Facebook users in the dataset, laying the groundwork for the article's personality-oriented research.

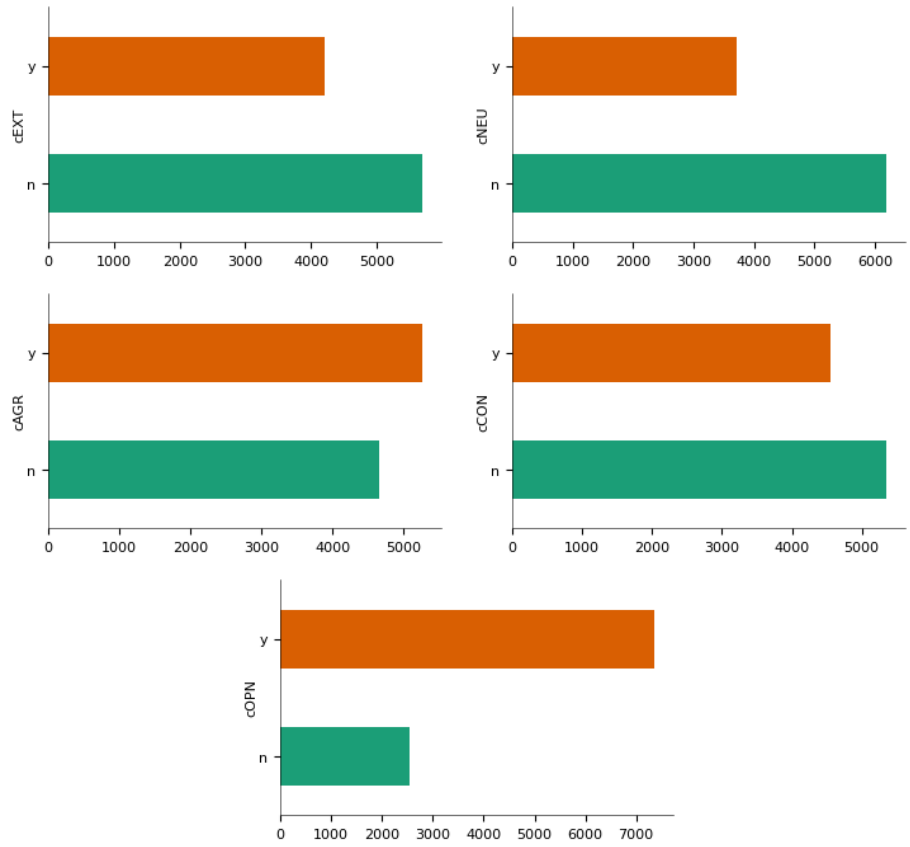

Figure 3.2: Visualization of Categorical Distribution of myPersonality dataset

Table 3.1: myPersonality Dataset Numerical Representation of Data Distribution

Trait	Total 'n'	Total 'y'
cEXT	5707	4210
cNEU	6200	3717
cAGR	4649	5268
cCON	5361	4556
cOPN	2547	7370

3.1.2 The essay dataset

The second dataset used in this study is the assay dataset, which serves as the established benchmark in this field [24]. Curated by Pennebaker and Laura King, this data set consists of a large corpus of text written by 2,467 people between 1997 and 2004. The texts in the essay dataset have been carefully categorized according to different dimensions of personality traits. Notably, these studies have been carefully tagged with their respective authors' appropriate personality traits, making the dataset well-suited for supervised learning applications. It is important to emphasize that the authors of the studies in this dataset are students of the American Psychological Association, which contributes to the contextual understanding of the origin of the dataset. Figure 3.3 and Table **??** depict visual and numerical representations of the Big Five personality traits distribution throughout the essay sample. These representations provide a complete study of how many records or individuals fit into the 'y' and 'n' categories of each personality attribute.

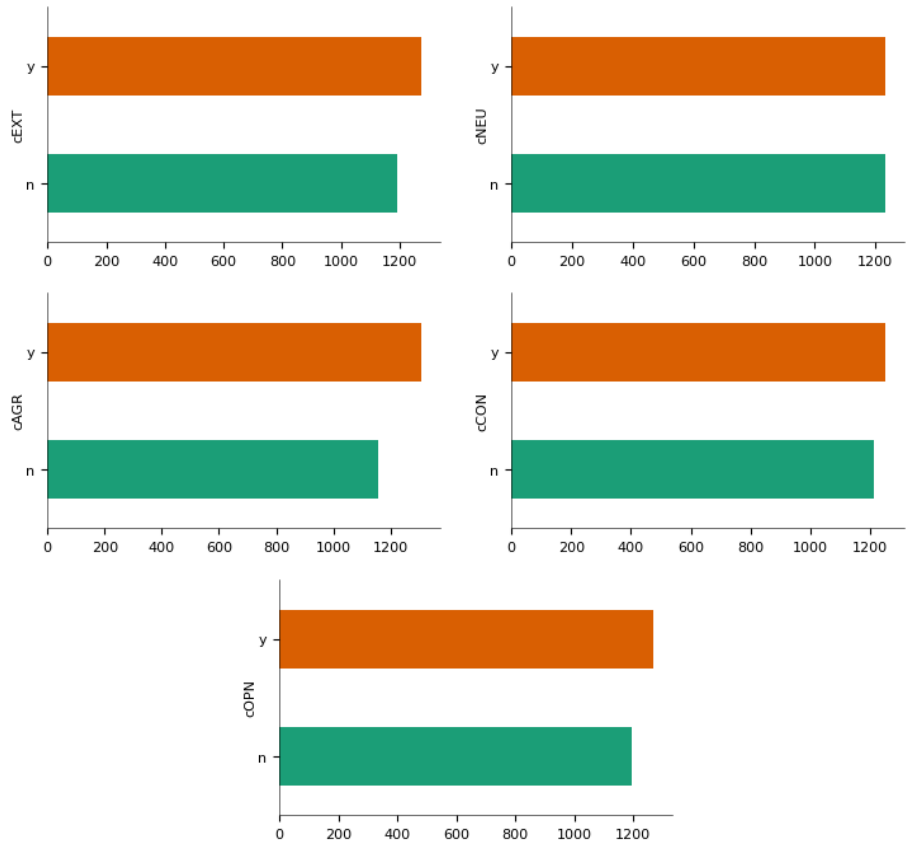

Figure 3.3: Visualization of Categorical Distribution of essay dataset

Table 3.2: Essay Dataset Numerical Representation of Data Distribution

Trait	Total 'n'	Total 'y'
cEXT	1191	1276
cNEU	1234	1233
cAGR	1157	1310
cCON	1214	1253
cOPN	1196	1271

3.2 Pre-processing

Pre-processing was performed on the "Essay" and "Facebook" datasets to prepare the textual data for in-depth analysis. The steps in this pre-processing phase were designed to enhance the textual material, making it more homogeneous, readable, and contextually relevant. Figure 3.4 represents the pre-processing stage of our methodology section.

The initial stage was Expand Contractions, such as "can't" and "I'm," were stretched to their full forms. This standardization provided text consistency and allowed for more accurate analysis. Hyperlinks, which are common in social media and web-based writing, were carefully deleted from the data. This stage was designed to filter out any potential noise or irrelevant information. It's worth noting that this phase was only applicable to datasets other than the "Essay" dataset, where hyperlinks were not permitted. To maintain data consistency and uniformity, all text was changed to lowercase. This change prevented the analysis from interpreting words with varied letter cases as distinct entities, resulting in more accurate results. Special letters, symbols, and punctuation marks that did not add substantial significance to the text were removed. This step increased the data's readability and homogeneity.

Tokenization is the process of separating text into its constituent elements to prepare it for further analysis. The text was tokenized into individual words or subwords after basic cleaning. Dates and timestamps, for example, were systematically eliminated from the text. As a result, rather than numerical data, the research concentrated only on linguistic patterns and textual content.

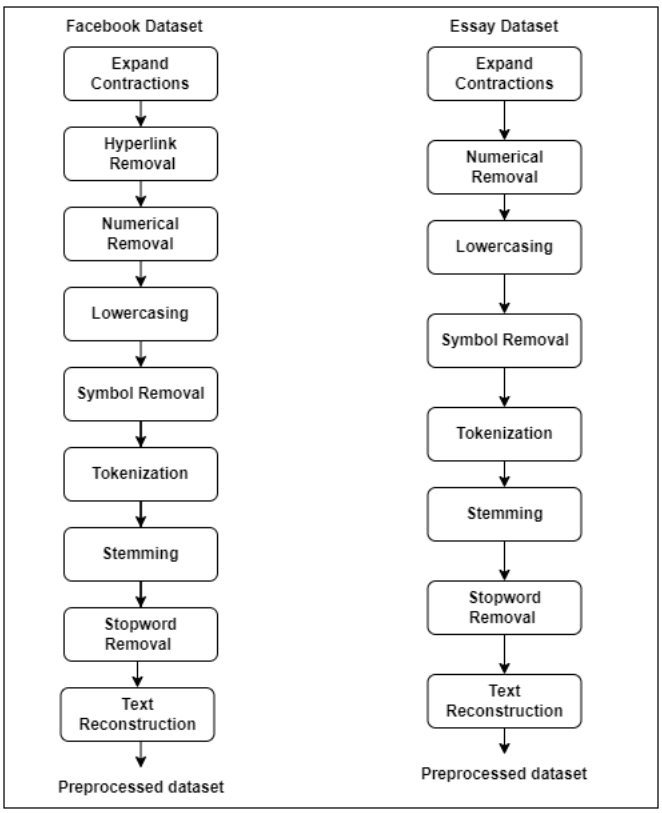

Figure 3.4: Pre-processing Stage

To normalize words to their base or root forms, word stemming was used. For example, the words "running" and "ran" were shortened to "run." This stage combined words with similar meanings, improving the accuracy of subsequent studies. to eliminate common stopwords in the English language. These stopwords, like "the" and "and," were omitted from the analysis since they frequently lack significant significance. A notable improvement in this phase was the decision to keep personal pronouns like "I," "you," and "they." Personal pronouns provide important context clues about the author's point of view and communication style.

While traditional stopwords were first removed, several common stopwords were kept to add context. This method attempted to create a balance between removing less informative stopwords and retaining personal pronouns for context. The cleaned text was reconstructed into understandable text strings after significant pre-processing operations. The result was material that was homogeneous, accessible, and contextually rich, laying the groundwork for more in-depth analyses in both the "Essay" and "Facebook" datasets.

These painstaking pre-processing efforts aided in increasing the quality of the

textual data, opening the way for later studies to uncover relevant patterns, feelings, and insights within the datasets.

3.3 Features Extraction

We extracted features from the datasets in two ways: utilizing NLP statistical approaches like word count, readability score, and so on, and using pre-trained transformers like Bigbird, Albert, and Distilbert.

3.3.1 NLP Statistical Feature

Analysis of personality traits from textual data involves the extraction and use of various statistical characteristics. These features provide valuable information about the various linguistic, emotional, and psychological dimensions inherent in the textual content.

In this study, we will use TF-IGM measurements instead of TF-IDF because study [10] was used in their study, resulting in higher precision. The use of TF-IGM in text analysis has shown improved performance in various applications compared to traditional methods such as TF-IDF (Term Frequency-Inverse Document Frequency). Researchers have found that TF-IGM can lead to higher accuracy and more informative features, making it a promising tool for discovering important linguistic patterns and associations in text data [3]. Sentiment analysis is about examining the subjectivity and dominant emotional tone in the text. Understanding emotional tendencies and patterns is instrumental in uncovering connections between emotions and various personality traits [4]. This analysis aims to determine the degree of subjectivity (whether the text is objective or subjective) and polarity (whether the sentiment is positive, negative, or neutral). Sentiment analysis provides researchers with information about a person's emotional state, opinions, and perspectives expressed through texts. The NRC lexicon understands words grouped according to their emotional connotation, with each word assigned affective labels corresponding to emotions such as fear, anger, confidence, joy, and more. Analysis of the lexical properties of NRC reveals the intricate relationships between emotions and personality traits and enriches understanding of linguistic expression and emotional disposition [1]. One of the functions of the NRC lexicon is counting the occurrences of these emotion words, making it possible to quantify different emotions present in a text. This approach provides information about a person's psychological makeup and emotional tendencies.

In addition to the key features described above, the analysis includes several linguistic and textual attributes that provide additional context and depth

to the personality assessment: Readability Scores: Metrics such as the Flesch Reading Ease and Gunning Fog Score measure the complexity and readability of text [2]. This highlights the ease with which different audiences can understand the content. The frequency of personal pronouns ("I", "you", "he", "she", "we" and "they") reveals self-referential tendencies and patterns of interpersonal communication. The attributes "word variety" and "average word length" reflect the richness and complexity of the vocabulary and indicate linguistic diversity and sophistication of expression. The social behavior count, a count of words related to social interactions and relationships, reveals a person's social behavior and provides information about possible associations with personality traits. Various count-based metrics, such as the number of capital letters, capital letters, repeated words, and the occurrence of proper nouns (PROPNAME), provide insight into different writing styles and patterns.

Table 3.3: NLP Statistical Features

Feature Name	Description	Feature Count
TF-IGM	Statistical method to find how important a word is in a document influenced by the class label of a document. This method is used based on the research performance comparison between TF-IDF and TF-IGM in text classification [12].	60
Sentiment Analysis	The sentiment analysis features include sentiment polarity, sentiment subjectivity, positive percentage, negative percentage, and neutral percentage. T the researcher used a polarity sentiment analysis approach [13] to extract these features.	5
Emotion-Based Features (NRC Lexicon)	Contains 14,000 sets of words in English and the relation of each word with eight common emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust [22]	8
Linguistic and Textual Attributes	This category encompasses various linguistic and textual attributes such as readability scores, pronoun usage (first-person, second-person, third-person pronouns), word diversity, average word length, word count, character count, and counts related to social behavior, capitalization, repeated words, and occurrences of proper nouns (PROPNAME)	17
Total Statistical Features		90

3.3.2 Pre-trained Feature Extraction

This research uses BigBird, ALBERT, and DistilBERT [14] [15] [16] to address the shortcomings of conventional models like BERT, RoBERTa, and XLNet by utilizing the distinct characteristics of each model. These models, each with a unique architecture that enables them to handle various linguistic patterns, were trained on enormous text collections. They can effectively catch the subtleties of violent language expressions thanks to their intrinsic systems,

which helps us grasp aggression in the text more precisely and insightfully.

The method of feature extraction entails converting the text into numerical representations that capture both the linguistic characteristics and the context of the words. Tokenization utilizing WordPiece tokenization, creation of input IDs, contextual embedding extraction using self-attention processes, pooling embedding creation, statistical feature incorporation, and feature vector formation for aggressiveness classification are all part of this process. WordPiece tokenization is used to tokenize the text, breaking it up into subwords or tokens. The input tensor is created by converting these tokens into input IDs.

Then, using self-attention mechanisms, the contextual embeddings—also known as self-attention embeddings—are extracted. The contextual embedding matrix is formed by these techniques, which allow each token to take into account its relationships with all other tokens. The contextual embeddings matrix is averaged to produce pooled embeddings, which offer a condensed representation of the meaning of the full text. To provide more context for the total feature vectors, linguistic attributes from the text are retrieved, including word diversity, pronoun usage, and others. We combine the pooled embeddings and statistical feature extraction to produce the embeddings. Each model uses this combined vector as a feature vector, which includes both language and contextual understanding characteristics. Using a thorough grasp of the text's context and linguistic features, these feature vectors are then applied to tasks requiring the classification of aggression.

Big Bird

BigBird is a transformative model in the realm of natural language processing (NLP), known for its exceptional ability to handle longer sequences of text. Trained on an extensive corpus of text data, it combines global and local attention mechanisms to capture both long-term dependencies and short-contextual information within the text. Global attention is a distinctive feature of BigBird, enabling it to comprehend the intricate linguistic patterns associated with complex topics such as aggressiveness. Simultaneously, local attention ensures that the model can grasp the finer details in the text.

BigBird's use of sparse attention patterns is a major advance. Traditional approaches, such as BERT, suffer from computational restrictions as the length of the sequence rises, rendering them unsuitable for exceedingly long texts. BigBird addresses this issue by attending to specific tokens selectively, considerably decreasing computing costs. This efficiency is especially useful when dealing with very long sentences, which conventional models struggle with.

Pre-training for BigBird is based on four publicly available datasets: Books, CC-News, Stories, and Wikipedia. It uses Roberta's sentence component vocabulary, which is taken from GPT-2. This extensive training data provides BigBird with a diverse set of linguistic patterns and global expertise. BigBird is distinguished by its capacity to handle sequences of up to 4096 tokens at a far lower computational cost than classic models such as BERT. It has shown cutting-edge performance on a variety of tasks demanding extremely long sequences, such as extended document summarization and question-answering with complex contexts.

BigBird's novel architecture, sparse attention mechanism, and extensive training in a variety of text sources make it a formidable model for processing and comprehending extraordinarily long text sequences, bringing substantial advances to the field of natural language processing.

Albert

ALBERT is a transformer-based model that excels at natural language processing. It was trained on a variety of text corpora, including English Wikipedia (2.8 billion words) and BooksCorpus (about 11,038 books). This large and diverse training dataset provides ALBERT with the capacity to comprehend a wide range of linguistic patterns and English subtleties.

The issue of parameter efficiency is one of the fundamental challenges that ALBERT addresses. Traditional models, such as BERT, sometimes include an excessive amount of parameters, necessitating large computing and memory resources. it devises ingenious techniques to bypass this constraint. During the pre-training phase, one such method is guessing what will happen next in a text. This exercise teaches ALBERT about the relationships between sentences and how language flows, allowing it to become skilled at interpreting context inside text. This contextual awareness is especially useful for tasks like recording aggressive language, which requires a comprehension of how distinct phrases connect. ALBERT also makes architectural improvements to increase parameter efficiency. It significantly reduces the overall number of parameters by employing factorization and parameter-sharing algorithms. This architecture optimization enables ALBERT to maintain good performance while employing fewer parameters, resulting in shorter training times without sacrificing efficacy.

it is a complex transformer model that excels at self-supervised language representation learning. It addresses the issues given by the computational needs of large-scale language models by leveraging diverse training data, unique pre-training procedures, and an efficient design, making it a useful tool in

natural language processing.

DistilBert

DistilBERT is a refined version of the BERT (Bidirectional Encoder Representations from Transformers) model that was trained using BooksCorpus and the English Wikipedia. It, like BERT, gains from exposure to a wide range of text sources, allowing it to understand and express a wide range of linguistic patterns. The fundamental issue that this model addresses is the size and resource requirements of models such as BERT. While BERT is powerful, its intricacy necessitates a significant amount of memory and computing capacity. By simplifying the design, decreasing the number of layers and parameters, and becoming substantially more resource-efficient, DistilBERT provides a solution. It is known for its efficiency. It is a viable option in situations where computational resources are limited or speed is critical. Surprisingly, despite its smaller size, DistilBERT preserves a significant amount of BERT's performance, making it an important tool for a variety of natural language processing jobs. It is particularly good at analyzing word groups to comprehend text. It combines several sorts of information, including word placement inside sentences and word relationships, to understand the meaning of text, including nuances and confrontational language.

It is a smaller, faster, and more resource-efficient version of BERT. Its ability to maintain significant performance while addressing the resource limits of its predecessor qualifies it for a wide range of natural language processing applications, particularly in cases where computational resources are limited or efficiency is a key priority.

3.4 Model Prediction

Our next goal after obtaining feature vectors from the previous steps was to classify these vectors into certain personality characteristic categories, namely EXT, NEU, AGR, CON, and OPN. We used two different recurrent neural network (RNN) designs to do this: Bidirectional Gated Recurrent Unit (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM).

The use of BiGRU (Bidirectional Gated Recurrent Unit) and BiLSTM (Bidirectional Long Short-Term Memory) recurrent neural network architectures for the classification phase of a text data processing task is based on their efficacy in dealing with sequential data. Text data is naturally sequential, with the arrangement of words and phrases frequently carrying substantial information. RNNs are a type of neural network that is specifically built to handle

sequential input, which makes them an excellent choice for text-processing jobs. BiGRU and BiLSTM are both bidirectional RNN variations. They can capture contextual information from both past and future elements in the sequence by processing input sequences in both forward and backward directions. Because word meanings frequently depend on their surrounding terms, bidirectional context modeling can be critical for interpreting the meaning of words in a phrase.

In the BiGRU architecture, we used bi-directional Gated Recurrent Units (GRUs). BiGRU's bidirectional nature allows it to analyze feature vectors both forward and backward. This bidirectional analysis improves the model's knowledge of the contextual relationships present in the feature vectors greatly. The forward and backward GRU outputs are concatenated before being processed by a linear layer, which gives predictions for the personality characteristic categories.

The BiLSTM model, on the other hand, includes Bidirectional Long Short-Term Memory units (BiLSTMs). LSTM units are well-known for their ability to capture long-term dependencies within sequential data, making them especially well-suited for applications involving complex temporal interactions. The BiLSTM model, like the BiGRU model, integrates the outputs of forward and backward LSTM units through a linear layer to predict personality traits.

We used cross-entropy loss because it is well-suited to measuring the dissimilarity between projected probability distributions and actual target distributions, the cross-entropy loss function is widely employed in machine learning, particularly in classification tasks. The purpose of classification is to assign each input to one of several classes, and the cross-entropy loss quantifies the amount of mismatch between the predicted probability given to each class and the true class labels. It promotes the model to give high probability to the right classes while penalizing deviations from the true distribution, making it a good choice for training classification models to optimize for accurate class predictions. Furthermore, it generates smooth gradients, allowing for efficient gradient-based optimization algorithms such as stochastic gradient descent, which are critical for training deep neural networks.

Parameter adjustment was performed to improve the model's performance. The rigorous hyperparameter tuning approach, which involves a grid search across many configurations, is common in machine learning. It is critical to discover the best settings for the model's performance. Hidden sizes, the number of layers, batch sizes, learning rates, and epochs are all important parameters in defining the model's capacity, convergence speed, and generalization ability. Grid search investigates several combinations to determine

the best arrangement for the task at hand. The power of the model to capture complicated patterns is affected by hidden sizes and the number of layers. Smaller values are more likely to underfit, while larger values are more likely to overfit. As a result, experimenting with different sizes is critical. The size of the batches affects the speed of training and convergence. Smaller batches may provide more frequent updates but are noisier, whereas larger batches may provide more steady updates but have a slower convergence. During optimization, learning rates dictate the step size. A high learning rate can lead to divergence, whereas a low rate can lead to slow convergence or becoming stuck in local minima. The number of training epochs is critical for the model to converge on the best solution. Too few epochs may result in underfitting, whereas too many epochs may result in overfitting.

3.5 Evaluation Matrix

Our classification models were thoroughly evaluated to ensure a thorough assessment of their performance. Key evaluation parameters such as accuracy and F1-score were used to assess the models' ability to predict personality traits. These criteria provided a comprehensive picture of the models' ability to appropriately categorize data. Each model was thoroughly evaluated for each personality attribute category. The evaluation procedure produced categorization reports, which were critical components of our review. These reports gave detailed insights into how the models performed across many aspects, allowing us to pinpoint specific areas that could benefit from improvement and refinement. Our evaluation process intended to provide a thorough picture of the strengths and shortcomings of our classification models by combining accuracy and F1-score measurements. This method guaranteed that our models were thoroughly tested, yielding useful insights for future improvements and optimizations in personality trait prediction.

3.6 Experiment

We used a multimodal strategy to feature extraction and personality trait classification in our experiments. The extraction of NLP statistical features was combined with the use of three sophisticated transformer-based models: BigBird, ALBERT, and DistilBERT. These models were pre-trained on large text datasets to obtain contextualized word and sentence representations, allowing them to capture nuanced language patterns.

We investigated two recurrent neural network (RNN) designs for the classification phase: BiGRU and BiLSTM. These RNN designs were chosen be-

cause of their shown competence in handling sequential data processing tasks, which corresponds to the nature of text data. We used a grid search to tune hyperparameters to find the best configuration for each model. This rigorous testing included altering hidden sizes (64, 128, and 256), the number of layers (2, 3, and 4), batch sizes (16 and 32), learning rates (0.001 and 0.0001), and epochs (5, 10, and 15).

We followed best practices for weight initialization in each model, ensuring that our models were ready for effective training. To load the training and validation sets effectively, DataLoader instances were used. We used the Cross-Entropy Loss, a reasonable solution for this type of problem, to address the multi-class categorization aspect of our work. We rigorously trained the model for each combination of hyperparameters during the training procedure, which lasted numerous epochs. The model with the highest validation accuracy was determined to be the most effective. We used major evaluation criteria, primarily Accuracy and F1-score to assess the performance of our models. These metrics gave a thorough evaluation of how successfully our algorithms classified personality traits. We wanted to discover the ideal configuration and architecture for personality trait categorization using state-of-the-art transformer-based models with NLP statistical characteristics through this broad experimental strategy.

CHAPTER 4

ANALYSIS & RESULTS

This section delves into the findings and conversations generated by our research on personality characteristic classification for the two datasets. Our research revolves around the use of cutting-edge transformer-based models, specifically BigBird, ALBERT, and DistilBERT. We also investigate the effect of introducing NLP statistical variables into these models. We give a thorough examination of our findings here, giving light on the performance, strengths, and opportunities for improvement seen across the various models and feature sets. We investigate how these models handle the challenging task of personality trait classification, as well as the complexities introduced by the addition of NLP statistical factors.

Our talks go into the ramifications of our findings, taking into account the larger context of personality trait analysis, natural language processing, and prospective applications of our research. We also investigate future research and enhancement opportunities to further develop the field of personality trait classification using cutting-edge transformer-based models and novel feature engineering. We hope to share helpful insights and critical reflections on our study findings in this part, adding to a better understanding of the interaction between advanced NLP models, statistical characteristics, and personality trait classification.

4.1 Facebook Dataset

4.1.1 BiLSTM model results

The BiLSTM model results shown in Table 4.1 shed insight into the performance of different transformer designs in predicting personality traits. The BigBird model had an accuracy of 56.69% and an F1-score of 64.33% for the Extraversion (EXT) trait. When the Albert model was used, the accuracy increased to 70.67% and the F1 score increased to 71.82%. The DistilBERT model, in particular, attained an accuracy of 78.62% and an F1-score of

79.07%. The addition of new NLP statistical variables greatly improved these models' predictive potential, with the BigBird+NLP and Albert+NLP models obtaining accuracies of 85.16% and 84.50%, respectively. These findings imply that transformer models may efficiently capture complex linguistic patterns related to extraversion and that the addition of additional factors refines their predictions even further.

The performance trend for the Neuroticism (NEU) trait was stable. The BigBird model had an accuracy of 70.42% and an F1-score of 37.74%, but the Albert and DistilBERT models had an accuracy of 76.04% and 87.39%, respectively, with F1-scores of 60.22% and 76.44%. The addition of NLP statistical features enhanced performance similarly to the extraversion trait, with accuracies of 85.97% and 86.78% for the BigBird+NLP and Albert+NLP models, respectively.

The improved results demonstrate the models' ability to capture language nuances related to neuroticism. A similar pattern was observed for the Agreeableness (AGR) attribute. Accuracy rates for the BigBird, Albert, and DistilBERT models were 84.04%, 83.28%, and 86.98%, respectively, with F1-scores of 18.18%, 12.23%, and 47.01%. The addition of NLP statistical features improved the models' performance again, with the BigBird+NLP model obtaining 92.35% accuracy and an F1-score of 74.62

The models performed admirably across the board for the Conscientiousness (CON) characteristic. The BigBird, Albert, and DistilBERT models achieved high accuracies of 95.39%, 98.02%, and 97.72%, with F1-scores of 26.02%, 79.79%, and 77.39%, respectively. With the addition of NLP statistical features, the BigBird+NLP and Albert+NLP models achieved accuracies of 98.48% and 97.52%, respectively.

Finally, the characteristic Openness to Experience (OPN) performed well in terms of prediction. The BigBird model had an accuracy of 96.40% and an F1-score of 32.38%, while the Albert and DistilBERT models had accuracies of 97.92% and 98.23%, respectively, with F1-scores of 69.63% and 81.68%. The addition of NLP statistical features had a minor impact on model performance, with the BigBird+NLP model attaining 98.33% accuracy and the Albert+NLP model achieving 97.21%.

Table 4.1: Facebook Dataset (BiLstm Results)

Traits	Metric	**Bigbird**	**Albert**	**Distilbert**	**Bigbird + NLP statistical features**	**Albert + NLP statistical features**	**Distilbert + NLP statistical features**
EXT	Accuracy	0.5669%	0.7067%	0.7862%	0.8516%	0.8450%	0.6084%
	F1-Score	0.6433	0.7182	0.7907	0.8277	0.8271	0.6529
NEU	Accuracy	0.7042%	0.7604%	0.8739%	0.8597%	0.8678%	0.7351%
	F1-Score	0.3774	0.6022	0.7644	0.7834	0.7599	0.5052
AGR	Accuracy	0.8404%	0.8328%	0.8698%	0.9235%	0.9063%	0.8323%
	F1-Score	0.1818	0.1223	0.4701	0.7462	0.7141	0.0461
CON	Accuracy	0.9539%	0.9802%	0.9772%	0.9848%	0.9752%	0.9463%
	F1-Score	0.2602	0.7979	0.7739	0.8454	0.8032	0.3977
OPN	**Accuracy**	0.9640%	0.9792%	0.9823%	0.9833%	0.9721%	0.9630%
	F1-Score	0.3238	0.6963	0.8168	0.8156	0.7179	0.4823

4.1.2 BiLSTM model results

The BiGRU model produced findings shown in Table 4.2 that differed from the BiLSTM design in some ways. The BigBird model had an accuracy of 62.41% and an F1-score of 62.83% for the Extraversion (EXT) trait. The Albert and DistilBERT models, on the other hand, displayed higher accuracy, with 77.96% and 68.59%, respectively. The addition of NLP statistical features had no discernible effect on the models' performance, with the BigBird+NLP model achieving an accuracy of 75.08The Neuroticism (NEU) trait showed similar patterns. The BigBird model had an accuracy of 70.97% and an F1-score of 51.40%, but the Albert and DistilBERT models had accuracies of 76.85% and 76.90%, respectively, with F1-scores of 69.51% and 51.90%. The addition of NLP statistical features resulted in moderate performance increases, with 76.19% accuracy for the BigBird+NLP model and 76.85% accuracy for the Albert+NLP model. The models' performance for the Agreeableness (AGR) trait was consistent once again. Accuracy rates for the BigBird, Albert, and DistilBERT models were 82.42%, 89.11%, and 85.46%, respectively, with F1-scores of 20.59%, 64.23%, and 52.40%. The addition of NLP statistical characteristics resulted in negligible performance changes, with the BigBird+NLP model reaching an accuracy of 87.18%.

The BiGRU model succeeded admirably in the example of Conscientiousness (CON). With F1 scores of 43.43%, 65.14%, and 68.26%, the BigBird, Albert, and DistilBERT models achieved excellent accuracies of 94.98%, 96.91%, and 97.32%, respectively. The addition of NLP statistical characteristics improved performance slightly, with the BigBird+NLP and Albert+NLP models reaching accuracies of 97.37% and 96.91%, respectively. The trait Openness to Experience (OPN) shows significant accuracy and F1-score values. The BigBird model had an F1-score of 54.66% and an accuracy of 96.30%, while the Albert and DistilBERT models had accuracies of 97.97% and 94.12%, respectively, with F1-scores of 76.47% and 56.06%. The inclusion of NLP statistical characteristics had a minor impact on model performance, with the BigBird+NLP and Albert+NLP models reaching accuracies of 98.23% and 97.97%, respectively.

Table 4.2: Facebook Dataset (BiGRU Results)

Traits	**Metric**	**Bigbird**	**Albert**	**Distilbert**	**Bigbird + NLP statistical features**	**Albert + NLP statistical features**	**Distilbert + NLP statistical features**
EXT	Accuracy	0.6241%	0.7796%	0.6859%	0.7508%	0.7796%	0.6859%
	F1-Score	0.6283	0.7106	0.6846	0.7185	0.7106	0.6846
NEU	Accuracy	0.7097%	0.7685%	0.7690%	0.7619%	0.7685%	0.7690%
	F1-Score	0.5140	0.6951	0.5190	0.6466	0.6951	0.5190
AGR	Accuracy	0.8242%	0.8911%	0.8546%	0.8718%	0.8911%	0.8546%
	F1-Score	0.2059	0.6423	0.5240	0.5125	0.6423	0.5240
CON	**Accuracy**	0.9498%	0.9691%	0.9732%	0.9737%	0.9691%	0.9732%
	F1-Score	0.4343	0.6514	0.6826	0.7451	0.6514	0.6826
OPN	Accuracy	0.9630%	0.9797%	0.9412%	0.9823%	0.9797%	0.9412%
	F1-Score	0.5466	0.7647	0.5606	0.7619	0.7647	0.5606

4.2 Essay dataset

The findings obtained from the essay dataset utilizing the BiLSTM and BiGRU models, along with BigBird, Albert, and DistilBERT embeddings, show a significant difference from the performance found in the myPersonality Facebook dataset. In this example, the prediction powers of the models were noticeably limited across all features. The Classification Reports show that most personality qualities have low precision, recall, and F1 scores. When assessing the performance of the BiLSTM model with the integration of BigBird and NLP statistical characteristics, for example, the findings show difficulties in discriminating qualities. While the accuracy measures do not show substantial accuracy values, the models struggle to identify each personality feature reliably. Similar trends may be seen in the BiGRU model findings with different embeddings. The poor performance can be attributed to a variety of variables, such as the unique nature of the essay dataset, potential noise or unpredictability in the data, and differences in writing styles and content compared to the myPersonality Facebook dataset. Furthermore, the addition of NLP statistical features did not result in significant gains, implying that the essay dataset's linguistic and structural qualities may not correspond well with the features used.

In the context of these results, it is critical to recognize the impact of dataset properties on model performance. For superior results, the essay dataset's heterogeneous content and language nuances may necessitate specialized preprocessing, model modification, or alternative architectures. These findings highlight the significance of dataset selection, feature engineering, and model selection for creating personality trait prediction models for various textual data sources. Further research, including data pretreatment approaches and model adjustments, could potentially improve these algorithms' prediction performance on essay datasets. While the algorithms performed well on the myPersonality Facebook dataset, the essay dataset's unique properties made effective personality trait prediction difficult. These findings illustrate the complexities of predicting personality traits from a variety of textual data sources, as well as the necessity for specialized techniques to handle dataset-specific peculiarities.

4.3 Comparison

Table 4.3 provides a complete assessment of previous research endeavors' personality characteristic outcomes using cutting-edge models. The table focuses solely on the myPersonality dataset, which includes several approaches

such as deep learning, machine learning, and model averaging. To assess the efficacy of their models, researchers used a mix of performance metrics, including f1-score and Accuracy. This collection offers a comprehensive view of the predictive capabilities of cutting-edge algorithms for personality trait inference within the myPersonality dataset, encompassing multiple algorithmic paradigms and evaluation criteria.

When these numbers are examined, a clear conclusion emerges: the advanced deep learning architecture is the clear winner, with improved model performance across the range of accuracy and f1-measure. When compared to all other techniques, this prominent position remains. Furthermore, the data reveal a compelling trend in which classifiers enriched with Natural Language Processing (NLP) characteristics outperform those depending only on individual pre-trained model features. This supports the idea that integrating NLP features leads to a significant improvement in model performance when forecasting personality traits.

Table 4.3: Comparison with previous research

Research	EXT	NEU	AGR	CON	OPN
Tandera et al. [11]	78.95% on MLP	79.49% on MLP	67.39% on CNN ID	62.00% on GRU	79.31% on MLP and CCN ID
M.Tadesse et al. [25]	78.6% On SNA+ XGB	68.0% On SNA+ XGB	65.3% On SNA+ XGB	69.8% On SNA+ XGB	73.3% On SNA+ XGB Also on LWIC+ XGB
Yuan et al. [16]	57.0% On CNN	60.0% On CNN	57.0% on CNN	58.0% on CNN	76.0% On CNN
Chowanda A. et al [22]	76.92% On Model Averaging	78.21% On Model Averaging	72.33% On XLNet + NLP Features	70.85% On Model Averaging	86.17% On Model Averaging
Our model	**85.16% on Big bird + NLP features + Bilstm**	**87.39% On Distilbert + Bilstm**	**92.35% On Big bird + NLP features + Bilstm**	**98.48% on Big bird + NLP features + Bilstm**	**98.33% on Big bird + NLP features + Bilstm**
Results based on F1-Score					
Zheng and Wu [26]	0.71 On PMC +LIWC + unigram	0.70 On PMC +LIWC + unigram	0.68 On PMC+ LIWC + unigram	0.64 On PMC +LIWC	0.65 On PMC+LIWC With or without “unigram”
Chowanda A. et al [22]	0.748 On Model Averaging	0.709 On XLNet + NLP Features	0.701 On XLNet + NLP Features	0.652 On Model Averaging	0.912 On Model Averaging
Our model	**0.82 on Big bird + NLP features + Bilstm**	**0.76 on Distilbert +Bilstm**	**0.74 on Big bird + NLP features + Bilstm**	**0.84 on Big bird + NLP features + Bilstm**	**0.81 on Big bird + NLP features + Bilstm**

4.4 Discussion

Our research is motivated by the need to improve the effectiveness of feature extraction for NLP techniques, particularly when it comes to identifying violent language expressions and analyzing personality traits from textual data. To solve the constraints of traditional models, we chose transformer-based models such as BigBird, ALBERT, and DistilBERT. These transformers have distinct designs built to handle a wide range of linguistic patterns, and they have been trained on large text sets, allowing them to catch details of language expressions with more precision. In terms of personality trait analysis, we chose the BiGRU and BiLSTM recurrent neural network designs due to their demonstrated success in dealing with sequential data, which is inherent in the text. The order of words and phrases in text often contains important information, and RNNs are designed to excel at processing such sequential input, making them an appealing candidate for our text-processing jobs. In essence, our technique selection is motivated by the need to improve the depth and accuracy of feature extraction in NLP by taking into account the unique qualities and capabilities of each chosen model or architecture to generate more insightful and precise outcomes.

The Facebook dataset was our primary focus, and models like BigBird, Albert, and DistilBERT performed admirably in grasping the rich linguistic clues linked with personality attributes. These models demonstrated their ability to encode and comprehend linguistic nuances, proving their capacity for trait prediction. The most important finding from our research is the significant improvement in predicting accuracy gained by using NLP statistical features. This emphasizes the importance of language context in accurately predicting personality traits. The success of transformer-based models in this context suggests that they are capable of navigating the extensive network of linguistic patterns indicative of multiple personality traits.

When we shift our attention to the essay dataset, though, an unexpected contrast appears. The models encountered a variety of difficulties, highlighting the difficulty of applying these models to varied textual data sources. Notably, the models struggled to achieve high accuracy and F1 scores, showing that the essay dataset's unique qualities, which included a wide range of topics and writing styles, offered formidable challenges. These findings underscore the importance of customizing preprocessing procedures and tailoring model architecture to the peculiarities of distinct datasets, emphasizing the complex nature of personality trait prediction in various textual data domains.

Our findings are consistent with broader trends in natural language processing (NLP) research, which emphasize the effectiveness of sophisticated deep

learning architectures when augmented by NLP characteristics. On the myPersonality dataset, in particular, our models outperformed established strategies, confirming the utility of combining transformer-based models with language features for enhanced trait prediction. This discovery highlights the current paradigm shift towards leveraging the capabilities of deep learning models for NLP tasks and personality trait prediction.

The relatively low performance observed for certain personality traits in the presented results is attributed to several underlying factors. First, the complexity and subtlety of these traits within textual data are posing significant challenges. Traits like EXT and NEU do not exhibit explicit linguistic markers, making it difficult for models to discern them accurately from text alone. The inherently context-dependent nature of personality traits can further complicate their prediction. In these cases, the training data doesn't encompass a diverse range of linguistic expressions for these traits, and models struggle to generalize effectively. Furthermore, certain personality traits are less frequently expressed.

The choice of model architecture also plays a significant role in the performance variations. Not all transformer models are equally effective at capturing the nuanced linguistic cues associated with different personality traits. While some traits may align well with the strengths of a particular model, others may not, resulting in lower performance. For instance, models like Bigbird and Albert, despite their capabilities, might not be optimized for traits that rely on intricate linguistic patterns or traits with less evident textual markers. In these cases, the model's architecture and inherent biases limit its ability to perform well on specific traits. It's imperative to consider the compatibility of the model's architecture with the characteristics of the data when aiming to improve performance on challenging personality traits.

CHAPTER 5

CONCLUSION & FUTURE WORK

We began on a voyage into the exciting domain of personality trait prediction from textual data in this comprehensive study, leveraging the capability of advanced deep learning models for feature extraction, notably BigBird, ALBERT, and DistilBERT. We also used NLP statistical features in combination with this transformer to improve the performance of models. BiGRU and BiLSTM, to classify five personality traits using Facebook and essay datasets. When combined with NLP statistical features and BiLSTM, BigBird achieves F1-scores of 0.82, 0.76, 0.74, 0.84, and 0.81 for the traits EXT, NEU, AGR, CON, and OPN, respectively, with accuracies of 85.16%, 87.39%, 92.35%, 98.48%, and 98.33% on the Facebook dataset. Our research involved two independent datasets, the myPersonality Facebook dataset, and an essay dataset.

Our models performed admirably in predicting personality traits from the rich tapestry of social media posts in the myPersonality Facebook dataset. The combination of context-rich embeddings and NLP statistical features resulted in considerable increases across key measures such as accuracy and f1-score. These findings highlight the value of using advanced deep-learning models to extract complex personality insights from the ever-changing landscape of online social interactions. The essay dataset, on the other hand, revealed a more complex landscape. Despite their expertise in the myPersonality environment, our models struggled with the difficult task of accurately predicting personality traits from the diverse and nuanced information inherent in essays. Accuracy and F1 scores all fell here, serving as a poignant reminder of the necessity for tailored procedures that smoothly correspond with the unique linguistic qualities of various textual sources.

Several paths call for research as we look ahead to future work. First and foremost, the refining of transformer-based models looks promising. Fine-tuning model architectures, experimenting with novel variations, and improving pre-training procedures could boost performance across a wide range of

textual contexts. Another frontier emerges feature engineering. We foresee integrating a greater range of linguistic and contextual data, such as sentiment analysis, grammatical properties, and topic modeling, in addition to contextual embeddings. This comprehensive technique has the potential to expand our understanding of personality expression within textual data beyond the limitations of word embeddings alone. Exploration of multimodal techniques beckons. The combination of textual data with non-textual modalities such as photos or audio has the potential to provide a more holistic and full view of people's personalities, paving the path for deeper insights. Cross-cultural studies appeal, with the goal of elucidating the interaction between culture and personality expression through textual data. The study of universality or cultural diversity in personality traits has the potential to improve our understanding of human behavior in a variety of sociocultural circumstances.

Finally, this study represents a substantial advancement in the field of personality trait prediction using textual data. As we map our way into the future, we do it with a greater understanding of the intricacies and opportunities that this area offers. By adopting these future work directions, researchers will be able to uncover new layers of insight, stimulate innovation, and continue pushing the frontiers of what is possible in the dynamic environment of personality prediction from textual data.

REFERENCES

[1] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (3) (2013) 436–465.

[2] R. F. Flesch, A new readability yardstick., The Journal of Applied Psychology 32 3 (1948) 221–33.
URL https://api.semanticscholar.org/CorpusID:39344661

[3] K. Chin, Z. Zhang, J. Long, H. Zhang, Turning from tf-idf to tf-igm for term weighting in text classification, Expert Systems with Applications 66 (09 2016). doi:10.1016/j.eswa.2016.09.009.

[4] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, Vol. 10, 2010.

[5] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, Advances in neural information processing systems 33 (2020) 17283–17297.

[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations (09 2019).

[7] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (10 2019).

[8] . K. M. Stillwell, D. J., mypersonality project website. (2015).
URL https://sites.google.com/michalkosinski.com/mypersonality

[9] J. Pennebaker, C. Chung, M. Ireland, A. Gonzales, R. Booth, The development and psychometric properties of liwc2007 (01 2007).

[10] X. Xue, J. Feng, X. Sun, Semantic-enhanced sequential modeling for personality trait recognition from texts, Applied Intelligence 51 (11 2021). doi:10.1007/s10489-021-02277-7.

[11] T. Tandera, Hendro, D. Suhartono, R. Wongso, Y. Prasetio, Personality prediction system from facebook users, Procedia Computer Science 116 (2017) 604–611. doi:10.1016/j.procs.2017.10.016.

[12] K. D. A. H. Prajapatd, Personality identification based on mbti dimensions using natural language processing, Vol. 8, 2022, pp. 1653–1657.
URL https://ijcrt.org/papers/IJCRT2006219.pdf

[13] P. William, A. Badholia, B. Patel, M. Nigam, Hybrid machine learning technique for personality classification from online text using hexaco model, in: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 253–259. doi:10.1109/ICSCDS53736.2022.9760970.

[14] A. Bruno, G. Singh, Personality traits prediction from text via machine learning, in: 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), 2022, pp. 588–594. doi:10.1109/AIC55036.2022.9848937.

[15] R. K. Cherukuru, A. Kumar, S. Srivastava, V. Kumar Verma, Prediction of personality trait using machine learning on online texts, in: 2022 International Conference for Advancement in Technology (ICONAT), 2022, pp. 1–8. doi:10.1109/ICONAT53423.2022.9725910.

[16] C. Yuan, J. Wu, H. Li, L. Wang, Personality recognition based on user generated content, in: 2018 15th International Conference on Service Systems and Service Management (ICSSSM), 2018, pp. 1–6. doi:10.1109/ICSSSM.2018.8465006.

[17] S. Bharadwaj, S. Sridhar, R. Choudhary, R. Srinath, Persona traits identification based on myers-briggs type indicator(mbti) - a text classification approach, in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1076–1082. doi:10.1109/ICACCI.2018.8554828.

[18] L. Zhou, Z. Zhang, L. Zhao, P. Yang, Attention-based bilstm models for personality recognition from user-generated content, Inf. Sci. 596 (C) (2022) 460–471. doi:10.1016/j.ins.2022.03.038.
URL https://doi.org/10.1016/j.ins.2022.03.038

[19] J. Yu, K. Markov, Deep learning based personality recognition from facebook status updates, in: 2017 IEEE 8th International Conference

on Awareness Science and Technology (iCAST), 2017, pp. 383–387. doi:10.1109/ICAwST.2017.8256484.

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

[21] S. H. . T. M. Mohades Deilami, F., Contextualized multidimensional personality recognition using combination of deep neural network and ensemble learning. (2022). doi:https://doi.org/10.1007/s11063-022-10787-9.

[22] S. D. C. A. e. a. Christian, H., Text-based personality prediction from multiple social media data sources using pre-trained language model and model averaging., in: J Big Data 8, 2021. doi:https://doi.org/10.1186/s40537-021-00459-1.

[23] K. El-Demerdash, R. A. El-Khoribi, M. A. Ismail Shoman, S. Abdou, Deep learning-based fusion strategies for personality prediction, Egyptian Informatics Journal 23 (1) (2022) 47–53. doi:https://doi.org/10.1016/j.eij.2021.05.004.
URL https://www.sciencedirect.com/science/article/pii/S1110866521000311

[24] J. Pennebaker, L. King, Linguistic styles: Language use as an individual difference, Journal of Personality and Social Psychology 77 (2000) 1296–312. doi:10.1037//0022-3514.77.6.1296.

[25] M. Tadesse, H. Lin, B. Xu, L. Yang, Personality predictions based on user behavior on the facebook social media platform, IEEE Access PP (2018) 1–1. doi:10.1109/ACCESS.2018.2876502.

[26] H. Zheng, C. Wu, Predicting personality using facebook status based on semi-supervised learning, in: Proceedings of the 2019 11th International Conference on Machine Learning and Computing, ICMLC '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 59–64. doi:10.1145/3318299.3318363.
URL https://doi.org/10.1145/3318299.3318363

Thesis Plag

ORIGINALITY REPORT

7%	5%	5%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	1%
2	d-scribes.philhist.unibas.ch Internet Source	1%
3	journalofbigdata.springeropen.com Internet Source	1%
4	Hussain Ahmad, Muhammad Usama Asghar, Muhammad Zubair Asghar, Aurangzeb Khan, Amir H. Mosavi. "A Hybrid Deep Learning Technique for Personality Trait Classification From Text", IEEE Access, 2021 Publication	<1%
5	digitalcollection.utem.edu.my Internet Source	<1%
6	u-aizu.ac.jp Internet Source	<1%
7	"Proceedings of Third International Conference on Computing, Communications, and Cyber-Security", Springer Science and Business Media LLC, 2023	<1%