北海道大学
HOKKAIDO UNIVERSITY

Language Media Laboratory
Research Group of Information Media Science and Technology
Division of Media and Network Technologies
Graduate School of Information Science and Technology
Hokkaido University

Dušan Radisavljević

A Study on Machine Learning-based Approaches for Personality Identification and Translation

A doctoral dissertation
supervised by
Prof. Kenji Araki

Sapporo, 2023

Thesis Abstract in English

The popularity of social media services has resulted in the increasing shift of our interactions into online spaces. This phenomenon and the rapid development of various chatting applications have led to textual messages driving most of our communications. Therefore, correctly interpreting the intentions and feelings that interlocutors try to convey through text has become increasingly more important.

While context and familiarity between people play a large part in adequately interpreting their communication, specific patterns they exhibit remain consistent over time. Personality psychology attributes these patterns and differences between people on an individual level to a concept called personality. Personality is the sum of individual differences present in behavioural, emotional, and cognitive patterns and remains relatively consistent over time and context. With this in mind, we can deduce that personality plays an essential part in communication, as it describes both individuality and consistency.

The significance of correctly understanding and interpreting an individual's personality has been the focus of many researchers. Recent decades saw efforts to leverage technological advancements and new computational algorithms for such a task, leading to the formation of the personality computing research field. While personality computing is still relatively new, there has been an increasing interest trend. However, due to the field's novelty, the lack of research standardisation in the evaluation criteria has made comparing works difficult. In addition, one more problem has been the lack of readily available data labelled with personality-relevant information. The main contributing factors are often the preference for different personality measurements and concerns over data privacy.

These two issues, most frequently cited as the main obstacles throughout many of the works on personality computing research, have been my primary motivation for researching the possibility of connecting different personality assessment methods. If it were possible to leverage a connection between different personality assessments successfully, it would effectively increase the readily available data for all the personality assessment methods involved. Additionally, taking an approach centred around understanding personality and its reflection in communication can contribute towards developing a standardised evaluation framework, within which it would be possible to successfully replicate and interpret differences in performance across different research works.

With this objective in mind, the study described in this thesis starts with the speaker identification task, seeking to establish the possibility of identifying interlocutors using only text transcripts of their communication. The novel transformer-based approach proposed in this part of the study has been proven able to predict to

whom the utterance belongs with a degree of certainty that outperforms the baseline approach, scoring above 70% on the F1 metric. In addition to this, the experiments have resulted in a large dataset based on the textual transcripts of the dialogues from a commercial video game, with over 70,000 utterances. To the best of my understanding, while previous efforts have examined the prospect of using fantasy texts for dialogue-related tasks, this is the first time data from a commercial video game has been collected and published with a purpose of being used for such a task.

Relying on the findings that answer whether or not textual communication reflects personal differences, the study further examines the exact reasons behind these differences – by looking into the relationship between text-based features and two different personality assessment models. The two models in question, namely the Big Five and the Myers-Briggs Type Indicator, have both shown a correlation with certain linguistic features when analysing text from the social media platform Reddit. While this finding confirms the linguistic properties often attributed to the Big Five model due to its lexical background, it has also offered novel insight into similar properties possibly being reflected in the Myers-Briggs Type Indicator, a less-researched personality model.

These findings were then employed later in the study to convert data from the more easily obtainable Myers-Briggs Type Indicator and another personality assessment model, the Enneagram, into the Big Five personality assessments. The detailed approach taken during this part of the experiment ensures the study's reproducibility and comparability. The result is a simple approach that has caused an increase of up to 13.2% in correlation strength on the per-measurement basis for the Pearson r correlation coefficient evaluation metric.

In order to better adapt the evaluation criteria to the properties of the domain as well as data, the best-performing approach for the translation of the Myers-Briggs Type Indicator and Enneagram assessments into the Big Five ones was then re-evaluated using the Spearman's rank correlation coefficient and a root mean squared error evaluation metrics. These re-evaluations have helped confirm the original findings and further substantiate the claim regarding which features and algorithm choice seem most effective for the task.

Thesis Abstract in Japanese

ソーシャルメディアサービスの普及により、私たちの交流はますますオンライン空間に移行している。この現象と様々なチャットアプリケーションの急速な発展により、私たちのコミュニケーションの大半はテキストメッセージで行われるようになっている。そのため、テキストを通じて対話者が伝えようとする意図や感情を正しく解釈することがますます重要となっている。

人々のコミュニケーションを適切に解釈するためには、文脈や親しみやすさが大きく影響しているが、人々が示す特定のパターンには一貫性がある。パーソナリティ心理学ではこのようなパターンや人同士の個人レベルでの違いを、パーソナリティと呼ばれる概念に帰着させている。パーソナリティとは、行動、感情、認知のパターンに存在する個人差の総体であり、時間や文脈にかかわらず比較的一貫したままである。このように考えると、個性と一貫性の両方を表すパーソナリティは、コミュニケーションにおいて不可欠な役割を担っていると推察される。

個人のパーソナリティを正しく理解し、解釈することの意義は、多くの研究者が注目してきたところである。ここ数十年、技術の進歩や新しい計算アルゴリズムの活用が進み、パーソナリティコンピューティングの研究分野が形成されるに至った。パーソナリティコンピューティングはまだ比較的新しい研究分野であるが、その関心はますます高まってきている。しかし、この分野の新規性から、評価基準に研究標準がないため、作品の比較は困難である。さらに、性格に関連する情報をラベル付けしたデータが容易に入手できないことも問題になっている。その主な要因は、異なる性格測定の好みとデータのプライバシーに対する懸念であることが多い。

この2つの問題は、パーソナリティコンピューティング研究に関する多くの研究において、主な障害として最も頻繁に挙げられており、私が異なる性格評価方法の接続の可能性を研究する最大の動機となっている。もし、異なる性格診断法をうまく接続することができれば、関係するすべての性格診断法の利用可能なデータを効果的に増やすことができる。さらに、パーソナリティの理解とコミュニケーションへの反映を中心としたアプローチをとることで、標準化された評価の枠組みを開発することに貢献し、その中で、異なる研究作品間でのパフォーマンスの違いをうまく再現し解釈することが可能となる。

この目的を念頭に置いて、本論文で説明される研究は、話者識別タスクから始まり、対話者のコミュニケーションのテキスト転写物のみを使用して対話者を識別する可能性を確立することを目的としている。この研究で提案された新しいトランスフォーマーベースのアプローチは、F1指標で70%以上のスコアを獲得し、ベースラインアプローチを凌ぐ確実性で、発話が誰のものかを予測できることが証明された。さらに、この実験では、商用ビデオゲームのダ

イアログのテキストトランスクリプトに基づく大規模なデータセットが得られ、70,000を超える発話があることがわかった。私の知る限り、ファンタジーテキストを対話関連タスクに利用する見込みは、これまでの取り組みで検討されてきたが、市販のビデオゲームのデータをこのようなタスクに利用するのは、今回が初めてのことである。

本研究では、テキストコミュニケーションに個人差があるか否かの答えが得られたことを踏まえ、さらに、テキストベースの特徴と2つの異なる性格評価モデルとの関係を調べることで、その違いを生み出す正確な理由を検証している。ビッグファイブとマイヤーズ・ブリッグス・タイプ・インディケーターという2つの性格診断モデルは、ソーシャルメディアプラットフォームRedditのテキストを分析したところ、いずれも特定の言語的特徴との相関が示された。この発見は、ビッグファイブの語彙的背景からビッグファイブの言語特性を確認するものであるが、同様の特性が、あまり研究されていないMyers-Briggs Type Indicatorの性格モデルにも反映されている可能性があるという新しい知見を提供するものである。

これらの知見は、より入手しやすいMyers-Briggs Type Indicatorと、もう一つの性格診断モデルであるエニアグラムのデータを、ビッグファイブの性格診断に変換するために、研究の後半で使用されている。実験のこの部分で取られた詳細なアプローチは、研究の再現性と比較可能性を保証するものである。その結果、シンプルなアプローチでありながら、ピアソンr相関係数の評価指標において、測定ごとの相関強度が最大13.2%増加した。

また、評価基準を領域やデータの特性に合わせるため、Myers-Briggs Type IndicatorとEnneagramをBig 5評価に変換する際に最も優れたアプローチを、スピアマンの順位相関係数と平均2乗誤差の評価指標を用いて再評価した。これらの再評価により、当初の知見が確認され、このタスクに最も効果的と思われる機能とアルゴリズムの選択に関する主張がさらに実証された。

Contents

1 Introduction **1**

1.1 Background and Motivation 1

1.2 Research Objectives 2

1.3 Novelty of the Research and its Contributions 3

1.4 Thesis Structure 4

2 Theoretical Background and Related Works **5**

2.1 Personality 5

2.2 History of Inquiry into Individual Differences 5

2.3 Personality Evaluation Tools 7

2.3.1 Big Five model 7

2.3.2 Myers-Briggs Type Indicator 8

2.3.3 Enneagram of Personality 9

2.4 Personality Computing 9

2.5 Related Work 11

3 Text-based Speaker Identification from Video Game Dialogues **13**

3.1 Introduction 14

3.2 Related Research in Speaker Identification 14

3.3 Data 15

3.3.1 Dragon Age: Origins Dialogue Dataset 15

3.3.2 LIGHT Research Platform Dataset 17

3.4 Methods 18

3.4.1 K-Nearest Neighbors 18

3.4.2 Convolutional Neural Network 18

3.4.3 Convolutional Neural Network with Utterance Concatenation 18

3.4.4 Bidirectional Encoder Representations from Transformers 19

3.5 Experiments . . . 19
3.6 Discussion . . . 20
3.6.1 Limitations of the Study . . . 22
3.7 Conclusion . . . 23

4 Analysing Personality in Textual Communication 25
4.1 Introduction . . . 26
4.2 Related Research in Personality and Language Usage . . . 27
4.3 Analysis and Results . . . 28
4.3.1 Big Five Analysis . . . 28
4.3.2 MBTI Analysis . . . 30
4.4 Conclusion . . . 32

5 From MBTI Types to Big Five Traits 35
5.1 Introduction . . . 36
5.2 Related Research in Personality Computing . . . 38
5.3 Methods . . . 40
5.3.1 PANDORA Dataset and the Baseline Approach . . . 40
5.3.2 Feature Selection Approach . . . 42
5.3.3 Model Selection Approach . . . 46
5.3.4 Ethical Approach to Personality Research . . . 50
5.4 Results . . . 50
5.4.1 Feature Analysis . . . 52
5.4.2 Model Selection . . . 62
5.5 Discussion . . . 64
5.5.1 Limitations . . . 66
5.6 Conclusion . . . 67

6 A Different Way of Evaluating Personality Recognition 69
6.1 Introduciton . . . 69
6.2 Re-Evaluation of the Previously Described Experiments . . . 70
6.3 Discussion and Concluding Remarks . . . 70

7 Conclusion of the Thesis 75
7.1 Thesis Summary . . . 75
7.2 Future Work . . . 77

Bibliography **91**

List of Publications **91**

Acknowledgements **93**

List of Figures

2.1 Sample Big Five personality test results 8
2.2 Sample MBTI personality test results 9
2.3 Enneagram illustration 10

3.1 Data distribution for the Dragon Age: Origins dataset 16
3.2 Data distribution for the LIGHT dataset 17
3.3 Confusion matrix of the speaker identification results achieved on the Dragon Age: Origins dataset 22

4.1 Relationship between MBTI and the Big Five model illustrated . . . 27

5.1 Google Books n-gram Viewer data for personality models 37
5.2 Google Trends data for personality models 38
5.3 Probability distribution of personality traits in PANDORA dataset . 42
5.4 Correlation heatmap of personality labels in PANDORA dataset . . . 43
5.5 Model stack illustration for transferring Big Five classification to regression 44
5.6 Personality prediction pipeline illustration 51

List of Tables

2.1 List of personality research works that use social media as a source of data . . . 12

3.1 Results of the speaker identification task . . . 20

4.1 Correlation of *Agreeableness* to LIWC categories . . . 29
4.2 Correlation of *Extroversion* to LIWC categories . . . 30
4.3 Correlation of *Neuroticism* to LIWC categories . . . 31
4.4 Correlation of *Introvert* type to LIWC categories . . . 32
4.5 Correlation of *iNtuitive* type to LIWC categories . . . 33
4.6 Correlation of *Feeling* type to LIWC categories . . . 34
4.7 Correlation of *Perceiving* type to LIWC categories . . . 34

5.1 Users and comments per personality model within the PANDORA dataset . . . 41
5.2 XGBoost hyperparameters overview . . . 49
5.3 Baseline results on the PANDORA dataset . . . 52
5.4 Pearson correlation coefficient between gold standard labels and predictions made by the baseline method . . . 52
5.5 Correlation between actual Big Five values and predicted Big Five classes using median split . . . 53
5.6 Pearson correlation coefficient between predictions made using the median split classes and gold standard values . . . 53
5.7 Correlation between actual Big Five values and predicted Big Five classes using quartile split . . . 53
5.8 Pearson correlation coefficient between predictions made using the quartile split classes and gold standard values . . . 54
5.9 Correlation between LIWC dimensions and MBTI types (1) . . . 55
5.10 Correlation between LIWC dimensions and MBTI types (2) . . . 56
5.11 Effectiveness of LIWC dimensions for personality prediction . . . 57
5.12 Personality prediction results using selected LIWC dimensions . . . 59

5.13 Personality trait prediction results without the Enneagram predictions in the feature set 60
5.14 Pearson correlation coefficient between Big Five traits and Enneagram predictions on the PANDORA dataset 61
5.15 Pearson correlation coefficient scores between personality traits and their predicted values that utilise subreddit participation 62
5.16 Personality prediction results on different regression algorithms 63

6.1 PANDORA Baseline evaluated with Spearman's rank correlation coefficient and RMSE 70
6.2 Regression models re-evaluated on the Spearman's rank correlation coefficient 71
6.3 Regression models re-evaluated on the RMSE metric 72

Chapter 1

Introduction

1.1 Background and Motivation

Communication is an essential part of one's everyday life. It is a primary way people interact with the world around them and how they understand it. The development of computer technologies during the last century has led to an interest in communicating with computers as well as efforts to make them behave more like people. This has prompted the creation of the first rule-based chatbot called ELIZA [1, 2]. While certainly ahead of its time, ELIZA did not use any artificial intelligence. Instead, it was given sentences from which it identified keywords and asked simple follow-up questions. Nowadays, years after the AI boom, which occurred in the early 2010s, there is an increasing interest in embedding machines and virtual assistants with human-like traits [3].

While Bickmore and Picard [4] have stressed the importance of implementing human-like emotional intelligence in machines, this remains a challenging task due to the complexity of human nature. The individual differences and reasons behind them have been a source of scientific inquiry for several millennia [5]. At the turn of the twentieth century, psychologists turned to the concept of **personality** to answer these inquiries.

The concept itself, however, is multifaceted and intrinsic and, as such, has been defined in various ways using different personality theories [6]. Additionally, personality is measured by a variety of tests that differ based on the field of interest. With the development of machine learning technologies, these personality models have subsequently seen increased interest in academic works. However, the difference in preference for specific models within certain fields, coupled with privacy concerns and issues with comparability between works in the research field [7] have made the procurement of reliable data difficult. While the popularity of social media services has helped alleviate some of these issues, the possible connection between measurements offered by different personality models, if leveraged, could help further increase the amount of data available for research. Consequently, there is a clear need to better understand and analyse the differences and relationships between measures used by these personality models – and, in the process, enhance our understanding of personality.

1.2 Research Objectives

The main research goal of this study is to develop a sustainable approach for increasing the amount of readily available, personality-relevant data. In doing so, the approach would lead to enrichment of the presently available corpora for personality research. Additionally, it would help utilise personality measurements for fields in which they would, otherwise, be considered unsuitable.

While research on personality that utilises modern computational algorithms has seen an increase of interest in the last couple of decades [5, 7], data scarcity induced by privacy concerns, as well as rapid developments in technology, have resulted in research mainly being focused on using different algorithms – rather than allowing for comparison with previous studies [7, 8, 9]. This lack of comparability, as well as usage of complex algorithms that rarely seek to leverage linear relationships, have, in turn, led the majority of research to answer the question of which novel technology is better rather than how and why it is able to tie certain features to the concept of personality. As a result, it is necessary to approach this objective from the perspective that furthers our understanding of personality.

Another goal is to research how exactly personality shapes textual communication, be it in the form of transcripts of conversations or actual interactions in online spaces. The following three research questions need answering to achieve this objective. The first question is how effective are systems at telling the difference between communication patterns used by different individuals? If they turn out to be effective, the second question is why were they able to do this, or in other words, in which way does personality affect communication patterns? The final question involves answering the first question from another direction – is it possible to use these communication patterns or specific words and phrases to identify one's personality?

The final research goal of this study is to approach the personality in an analytical way, in order to reveal novel insights. While the present research in personality successfully describes the concept, it only does so through the lens of a single personality assessment model. In order to better understand the concept itself, it is necessary to connect multiple models through common features. As such, this goal, despite being more general in nature than the two previously described, should be fulfilled through their methods.

Therefore, in this thesis, I describe a three-step approach which seeks to:

- Test if textual communication patterns can be used to distinguish between individuals.
- If so, answer why, or in other words, what features of the language contribute to this.
- Test this connection between differences in individuality and language across different personality assessment models, seeking to leverage their similarities to convert from one into another.

1.3 Novelty of the Research and its Contributions

The research described in this work has resulted in several contributions while seeking to fulfil the previously mentioned research objectives.

First, I propose a simple yet effective transformer-based solution for the speaker identification task that only uses text data. While the use of text features in the research field has been previously examined, in recent years, it has been neglected in favour of acoustic and visual ones. As such, the application of modern deep learning approaches and their capabilities of predicting interlocutors were never before tested on text data. The approach resulting from my work has significantly outperformed the baseline methods, leading to a score of over 70% on the F1 metric when identifying individual speakers using only their utterances in text format.

My research in speaker identification has additionally resulted in a large-scale dialogue dataset consisting of over $70,000$ utterances originating from a commercial video game. To the best of my knowledge, this dataset is the first of its kind to be centered solely around dialogue, and as such has been used to evaluate the proposed transformer-based approach.

While the novelty of this research in examining the possibility of using commercial video games as a source of data for the speaker identification task has opened the possibility of using different interactive storytelling mediums for similar purposes in the future, the main strength lies in the generic nature of the method proposed. Due to it, the transformer-based method, and thus the contribution of my research, are transferable to different data sources, even real-life conversation transcripts.

Second, I have analysed the linguistic relationship between different personality models and psycholinguistic features present in text messages originating from a social media platform. This research, which focused on two different personality evaluation models – namely the Big Five and the Myers-Briggs Type Indicator – has further substantiated the findings of previous research works that noted the linguistic properties of the Big Five model. However, the main novelty of this research is its ability to prove similar properties for the Myers-Briggs Type Indicator. While research in the past has extensively covered this aspect of the Big Five model, due to the criticism of the Myers-Briggs Type Indicator, existence of similar properties for it were not examined. The importance of this finding lies in the fact that it not only further confirms the relationship between the dimensions used by these two models, which was noted in past research works (e.g. the work of McCrae and Costa [10]), but also sheds light on the nature of their connection.

Third, my research in personality computing and experimentation in translating the Myers-Briggs Type Indicator and Enneagram dimensions into the more scientifically-backed Big Five ones has resulted in a new approach that leverages the linear relationship between these personality evaluation models. The approach in question further builds on my analysis of the linguistic properties of the Myers-Briggs Type Indicator by using a combination of n-grams and different psycholinguistic features selected due to their relationship with the personality evaluation model. These features were then evaluated on different regression algorithms, with an approach that uses a combination of L^1 and L^2 normalisation showing the best results. This approach has achieved an increase of between 0.012 and 0.033 points

on the Pearson correlation coefficient metric, or up to a 13.2% increase in correlation strength for the predicted and actual values of the Big Five personality traits on the per measurement level.

Additionally, the detailed experimentation and comprehensive reporting style adopted while experimenting with the effectiveness of different features and algorithms helps better understand their effectiveness in personality recognition. A further benefit of this approach is that it allows direct comparison with different previous works on a more detailed level, as issues with comparability have been noted by different researchers in the past to be prevalent in the field of personality computing [7]. For this purpose, the approach was further evaluated using two additional metrics – Spearman's rank correlation coefficient and root mean squared error, as the usage of more than a single evaluation metric has been suggested by authors in the past [9].

Finally, the main novelty of this dissertation as a whole lies in the cohesive narrative and its ability to connect different research fields, namely personality computing and speaker identification. By doing so, this work helps further highlight how personality influences differences in communication styles while also proving that deep learning models can identify individuals by solely using the way these differences are reflected in textual data.

1.4 Thesis Structure

Besides the Introduction, this thesis consists of six other chapters. Chapter 2 provides theoretical background in psychology and the history of personality computing, as to better contextualise my motivation, design choices, and research in general. The chapter is theoretical, although very brief. As such, it is not intended as a substitution for an actual course in psychology but rather a short overview of the research field.

Chapter 3 details my endeavours in researching speaker identification task on textual data originating from a fantasy dialogue transcripts. Chapter 4 seeks to further expand upon the results of the preceding chapter by providing answers to the relationship between language usage and personality. It analyses the correlation between certain linguistic features and two different personality assessment models. Chapter 5 focuses on the issues that motivated this thesis and describes my efforts to translate the more easily obtainable personality measures into those that belong to a more scientifically relevant personality assessment model. As Chapters 3 to 5 address distinct research goals, problems, and domains, the corresponding works relevant to each chapter's research field are described separately within each chapter.

Chapter 6 reflects on the limitations of the study described in Chapter 5 from the perspective of issues the research field is facing and details my experiment using a different set of evaluation metrics. Finally, Chapter 7 concludes the thesis and provides insight into possible future directions for the work.

Chapter 2

Theoretical Background and Related Works

The purpose of this chapter is to provide a gentle introduction to the research area and the concept of personality in general. I introduce the concept briefly and outline its importance in daily life. After this, I briefly summarise its history and development in the fields of philosophy and psychology. I follow this with a brief discussion of widespread personality measurements often used in practice to quantify and qualify it. The penultimate section covers personality computing, a research field to which the experiments described in this thesis primarily belong. Finally, the last section includes some of the related works that conclude this chapter.

2.1 Personality

The concept of personality has emerged from the field of personality psychology to explain the differences present on an individual level between people. Due to the complexity of the topic it seeks to explain and its intricate nature, there is no single agreed-upon definition, with interpretation depending on the personality theory used to describe it. However, it can briefly be summarized as a set of distinct emotional, cognitive and behavioural patterns that differ from individual to individual but remain relatively consistent over time and context. As such, personality plays a pivotal role in determining one's life and identity, describing consistent behaviours and explaining that person's individuality. It influences various life choices individuals make [5], how they are perceived by others [11, 12] and also has a significant impact on how they experience the world around them [13].

2.2 History of Inquiry into Individual Differences

Personality psychology originated in the early 20th century; however, it is speculated that the efforts to classify people based on the communication, thinking and behavioural patterns they exhibit long predates written sources [5]. *Characters* [14] by the Greek philosopher Theophrastus is the earliest known literary work that touches upon individual differences. Dating back to the fourth century BC, it

includes 30 short descriptions of different moral types, known as characters, that can be interpreted as prototypes of the modern personality types. Some translators of the work have since noted that the word "trait", rather than "character", would be better suited, as certain characteristics overlap between the descriptions [15].

Another early work that has proven crucial for developing modern personality theories is that of physicians Hippocrates and Galen of Pergamon, which was later documented in Galen's book *De Temperamentis* [16]. Despite appearing roughly five centuries after *Characters*, their work on the Four Humours theory has arguably had a more significant influence on modern personality psychology and philosophy.

Hippocrates was the first to suggest that an imbalance in humour, or vital bodily fluids (from the Latin *humor* – meaning fluid), can influence behaviour. He described each humour as a combination of values assigned to two pillars – dry/wet or hot/cold. For instance, blood was hot and wet, while black bile was a result of combining the cold and dry pillars. Following their work, Galen speculated the existence of a moderate value between the two pillars. He combined the values into nine different temperaments, four of which he called primary [17]. These temperaments – namely, *sanguine*, *choleric*, *melancholic* and *phlegmatic* – have impacted both the English language and various works involving personality.

Philosophers and psychologists have thoroughly explored the theory of the four temperaments formulated by Hippocrates and Galen, attempting to explain the reasons behind individual differences. Prominent philosopher Immanuel Kant further explored the theory in his book *Anthropology From a Pragmatic Point of View* [18], arguing that, rather than nine, there are, in fact, only four temperaments. He described these temperaments as independent from one another and formulated a comprehensive list of traits to describe them. Some modern personality models, namely those that focus on different personality types, function by using a similar approach.

In addition to inspiring Kant's work, the four temperaments theory has also drawn interest from Wilhelm Wundt, widely considered the father of experimental psychology. Wundt proposed that a two-dimensional approach was sufficient to describe personality accurately. He introduced the dimensions of emotional intensity (strong – weak) and activity changeability (changeable – unchangeable) and expressed the four temperaments along the axes of these dimensions. Some authors have since pointed out that these interpretations of the four temperaments made by Kant and Wundt have an "uncanny" resemblance to *Neuroticism* and *Extroversion* – two dimensions that belong to the Big Five personality model [19].

The seminal contributions made by Immanuel Kant and Wilhelm Wundt can be seen as highly influential in the field of personality research [20], as they have helped influence several theories of prominent researchers. Some of the most notable examples include Sigmund Freud, Carl Gustav Jung and Gordon Allport, all well-respected in their research fields.

Austrian neurologist and founder of psychoanalysis, Sigmund Freud, is widely considered to be one of the most influential figures in the history of psychology. The theoretical underpinnings outlined in his psychoanalytical theory have revolutionised the field and influenced our understanding of the human mind. His work contributed towards a broader discussion about personality by linking inborn temperaments and

early experiences to subsequent behaviours throughout life [21]. Freud's research, while not directly influenced by the contributions of Wilhelm Wundt, can largely be seen as an outgrowth of Wundt's study of consciousness and behaviour. Similarly, while his research differs from that of Immanuel Kant, it is Kant's conceptualisation of morality and its effect on human behaviour that have served as foundational influences on Freud's theories [22].

Similarly to Freud, the traces of research done by Kant and Wundt can be found in the theories that Carl Gustav Jung and Gordon Allport have put forward. For example, the concepts of introversion and extraversion that Jung proposed trace back to his study of Kant's work on morality. In the case of Gordon Allport, his theory on personality traits that emphasises the differences present on an individual level can be viewed as a direct continuation of Wundt's research. Additionally, the concept of "cardinal trait" that is also present in Allport's work, which refers to a single dominant trait that shapes the personality, can be tied to the Kantian idea of self.

2.3 Personality Evaluation Tools

Throughout the history of personality psychology, several prominent theories and models have been used to describe personality [6, 23, 24, 25]. However, for the sake of conciseness, in this section, I will focus only on the three personality models that are relevant to this research and which are most commonly used in practice. These models are the Big Five model, the Myers-Briggs Type Indicator (MBTI) and the Enneagram of Personality.

2.3.1 Big Five model

The Big Five model or the Five Factor Model [26] has often been described as the "dominant paradigm" in the field of personality research and as "one of the most influential models" in psychology [27]. Its origins can be traced to a list of $4,500$ terms relating to personality traits introduced by Gordon Allport and Henry Odbert [28]. This list was initially reduced through factor analysis to 16 traits, only to be narrowed down to the final five from which the model received its name. As such, it results from contributions from several authors, with roots in the English lexicon [29]. The five traits or factors that make up the Big Five model are most frequently labelled as:

1. *Openness* – measure of curiosity;
2. *Conscientiousness* – measure of efficiency;
3. *Extroversion* – measure of energy;
4. *Agreeableness* – measure of compassion;
5. *Neuroticism* – measure of sensitivity.

When using the Big Five model to measure personality, each person is assigned a continuous value for each trait (Figure 2.1) with the exact scale of these numbers mainly depending on the test used for measurements [30]. An example of this would be a person scoring 85/100 in the *Openness* trait, which indicates they are less likely to be cautious when exploring new things.

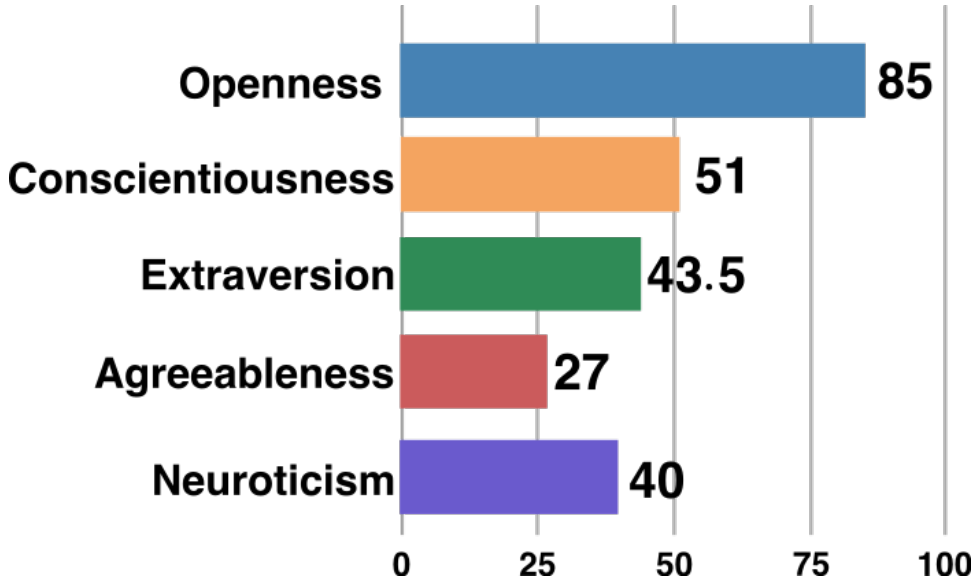

Figure 2.1: Graphical depiction of a sample Big Five test result.

2.3.2 Myers-Briggs Type Indicator

The *Myers-Briggs Type Indicator* [31], or the *MBTI* for short, is another prevalent personality evaluation tool. The Myers-Briggs Company[1], which is in charge of distributing the evaluation test, has stated that millions of people use the test annually in over 100 countries.

Unlike the Big Five model, MBTI focuses on personality types rather than traits and, as such, is based on the theoretical works of Carl Gustav Jung [32], who is often credited as the progenitor of personality types. Building on the three dichotomies that Jung originally proposed, namely *Introverted/Extroverted*, *Sensing/Intuitive* and *Thinking/Feeling*, Isabel Briggs-Myers and Katharine Cook Briggs later introduced the fourth dichotomy labelled as *Judging/Perceiving*, which finalized the initial MBTI model [33]. These dichotomies explain the following measures:

1. *Extroverted/Introverted* – Describes how an individual gains energy. Often abbreviated as *E-I*;
2. *Sensing/Intuitive* – Describes how an individual gains information. Often abbreviated as *S-N*;
3. *Thinking/Feeling* – Describes how an individual makes decisions. Often abbreviated as *T-F*;

[1] https://www.themyersbriggs.com/en-US/Connect-with-us/Blog/2018/October/MBTI-Facts--Common-Criticisms; last accessed on the 26th of May 2023

4. *Judging/Perceiving* – Describes how an individual observes the world around them. Often abbreviated as *J-P*.

Personality measured by the MBTI model is based on the idea that every individual has one pronounced value from the four dichotomies mentioned (Figure 2.2). For example, the *INTJ* type would refer to an individual who is *Introverted*, *iNtuitive*, *Thinking* and *Judging*. Thus, the MBTI offers a total of 16 different combinations that describe 16 unique personality types [34].

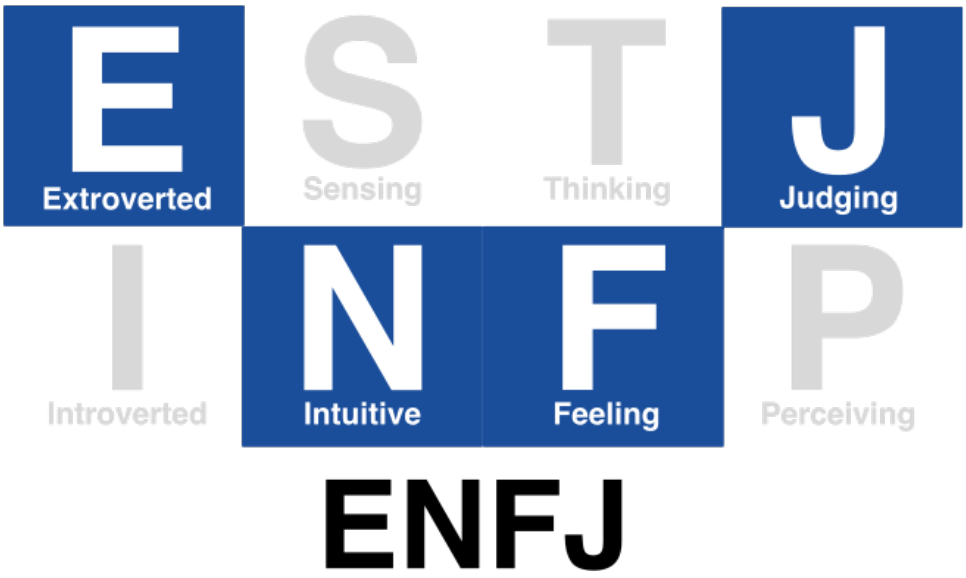

Figure 2.2: Example MBTI test results. The blue background indicates the dominant value of a dichotomy.

2.3.3 Enneagram of Personality

The *Enneagram of Personality* is another prominent personality model that uses types to measure individual differences [35]. The origins of the MBTI and the Big Five model can be traced to some of the early philosophical theories on individuality [20]; however, the exact origins of the Enneagram are disputed, with the Armenian philosopher George Gurdjieff often being credited with introducing it to the Western world [36].

The Enneagram is usually depicted as a circle with nine equidistant points connected with intersecting lines – the figure from which the model received its name. Personality types offered by the Enneagram are referred to by a number from *One* to *Nine* (e.g., Type *Eight*). Each number, or rather – type, is associated with different virtues, vices or ego fixations (depicted in Figure 2.3).

2.4 Personality Computing

In recent years, several studies have pointed towards the existence of a relationship between personality and essential aspects of life, such as career selection

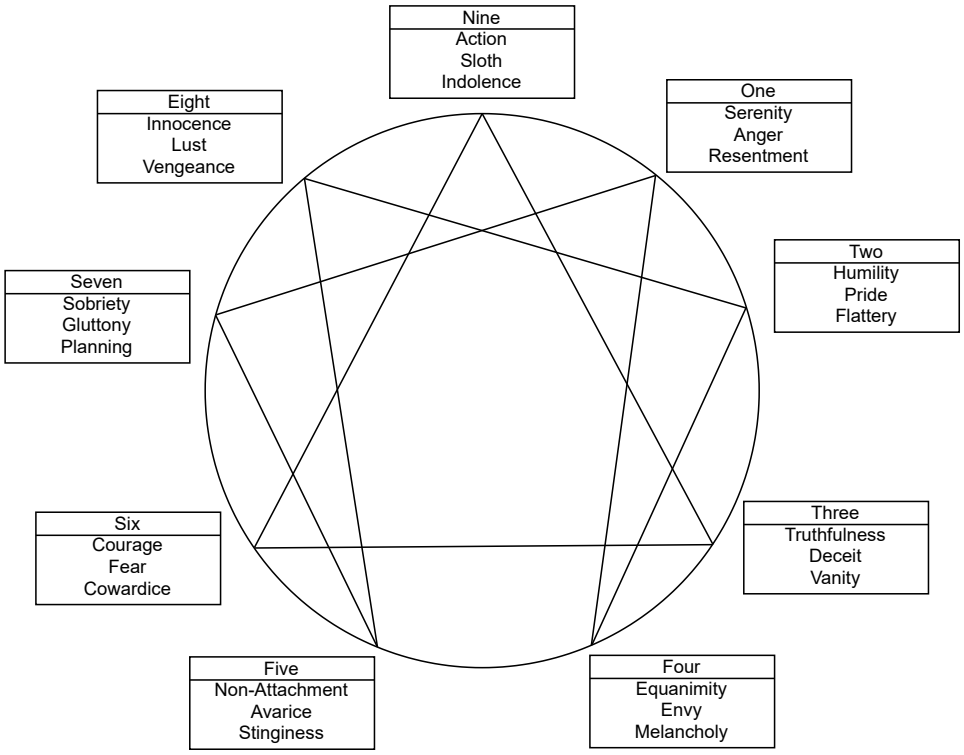

Figure 2.3: Enneagram types with their virtues, passions and ego-fixations as described by the Bolivian philosopher Óscar Ichazo.

and success [37, 38, 39], political participation and affiliation [40, 41, 42], religion [43], investment in social roles [44] and quality of life [45]. Due to the importance of personality and the rapid development of new technologies, Vinciarelli et al. [5] coined the term *personality computing* to describe the practice of applying machine-learning approaches to tasks involving personality [7].

The same authors described the main tasks of personality computing to be:

1. *Automatic Personality Recognition* – the identification of one's true personality from verbal or behavioural evidence.
2. *Automatic Personality Perception* – the recognition of the personality that others might assign to an individual, based on that individual's behaviour.
3. *Automatic Personality Synthesis* – the generation of artificial personalities.

With the research on personality traditionally using both the self-reported personality assessments and the scores assigned to one's personality by close friends or acquaintances, the first two tasks can be seen as an automatic approach to obtaining personality measures. The third task can be described as an effort to apply modern technologies to enrich the amount of personality-relevant information available for various studies.

Vinciarelli et al. concluded that any study that seeks to understand, predict or synthesise human behaviour could greatly benefit from personality computing approaches, closely relating the research field to affective computing [46] and its endeavours to build emotionally intelligent machines.

2.5 Related Work

It is difficult to exactly place the study of this thesis into a single research field, as it covers different tasks across its chapters. However, considering the objectives of the thesis, which include the improvement and development of a sustainable framework for the conversion from a set of measurements of one personality assessment model into another, I can best classify it as belonging to the research field of personality computing. While every chapter that follows will discuss works directly related to the task described within it, the following passage provides a more general look into the personality computing, and works that helped develop it as a research field.

Personality computing developed thanks to the early research works that involved personality, which were all centred around smaller sources of data – such as essays [47, 48, 49], emails [50] or blogs [51, 52]. An important turning point for the research field was the release of the *MyPersonality* dataset, which utilised information from the social media platform *Facebook*[2]. The dataset originated from the research of Kosinski et al. [53] and consisted of about 15.5 million Facebook statuses and some 7.5 million user profiles. As such, it represented the first publicly available large-scale dataset that included labels for the Big Five model. While *MyPersonality* has since been removed from the internet due to privacy concerns, with only a small portion of it being left accessible [54], it inspired many future researchers to examine social media platforms as a potential source of data (non-exhaustive overview provided in Table 2.1) [7, 8, 55].

[2] https://www.facebook.com/; last accessed on the 26th of May 2023

Table 2.1: A non-exhaustive overview of various personality computing works that utilise social media platforms as a source of data. Majority of the works listed tend to focus solely on the Big Five model.

Origin	**Authors**	**Modality**	**Personality Model**
Facebook	Schwartz et al. [56]	Text	Big Five
	Farnadi et al. [57]	Text	Big Five
	Verhoeven et al. [58]	Text	Big Five
	Celli et al. [59]	Text	Big Five
	Park et al. [60]	Text	Big Five
	Youyou et al. [61]	Text	Big Five
	Segalin et al. [62]	Images	Big Five
	Tandera et al. [63]	Multimodal	Big Five
	Kulkarni et al. [64]	Text	Big Five
	Ramos et al. [65]	Text	Big Five
	Xue et al. [66]	Text	Big Five
	Marengo et al. [67]	Text	Big Five
Flickr	Cristani et al. [68]	Images	Big Five
Instagram	Osterholz et al. [69]	Images	Big Five
Reddit	Gjurković and Šnajder[70]	Text	MBTI
	Wu et al. [71]	Text	MBTI
	Gjurković et al. [30]	Text	Big Five, Enne., MBTI
	Radisavljević et al. [72]	Text	Big Five, MBTI
Sina Weibo	Zhou et al. [73]	Text	Big Five (Extroversion)
TikTok	Meng and Leung [74]	Multimodal	Big Five
Twitter	Plank and Hovy [75]	Text	MBTI
	Verhoeven et al. [76]	Text	MBTI
	Tighe and Cheng [77]	Text	Big Five
	Celli and Lepri [78]	Text	Big Five, MBTI
	Balakrishnan et al. [79]	Text	Big Five, Dark Triad [1]
	Cahyani and Faishal [80]	Text	Big Five
Youtube	Biel and Gatica–Perez [81]	Video	Big Five
	Bassignana et al. [82]	Text	MBTI

[1] Dark Triad is a group of three traits associated with negative behaviour. These traits are largely independent from the traits measured by the Big Five model.

Chapter 3

Text-based Speaker Identification from Video Game Dialogues

In this chapter, I detail the investigation into the possibility of using textual data originating from fantasy dialogues for speaker identification. While most of the existing work focuses on using acoustic features for the task, here I outline the transformer-based machine learning approach that utilises text-dependant features only. In addition to improving the existing work in the field, which was used as a baseline, this research also introduces a new dataset acquired from a commercial video game. To the best of my knowledge, this is the first time this task has been conducted in the domain of video game dialogues.

The main contributions of this chapter are summarised as follows:

1. The proposal of a simple yet effective transformer-based approach to speaker identification that only utilises textual data.
2. To the best of my knowledge, this work is the first of its kind to examine the possibility of using video games as a data source for the speaker identification task. As such, it further opens the possibility of using interactive storytelling mediums for the said task.
3. While there have been previous endeavours into creating video game-related corpora for an NLP task using a commercial video game (e.g. the work of Bergsma et al. [83]), the dataset introduced as part of this chapter is the first of its kind to be centred around video game dialogues. The entire dataset is accessible from the following GitHub URL: https://github.com/dradisavljevic/DAODataset.
4. The approach detailed in this chapter is designed to be general, irrespective of the data; thus, it is transferable to transcripts of real-life dialogues.

The following peer-reviewed publication has served as the basis for this chapter:

- Radisavljević, Dušan, Bojan Batalo, Rafal Rzepka, and Kenji Araki. "Text-Based Speaker Identification for Video Game Dialogues." In Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3, pp. 44-54. Springer International Publishing, 2022. [84]

3.1 Introduction

Speaker identification is the task of connecting an utterance to the speaker that produced it. It is helpful in developing dialogue systems, with some authors citing it as a pivotal step in developing believable human agents [85]. Given a set of possible speakers, the problem of speaker identification can be formulated as a simple classification problem: which utterance belongs to which class, or rather, to which speaker.

In their daily lives, people unconsciously identify speakers by gathering audiovisual information from their utterances, such as the direction from which the sound is coming or even vocal timbre. Nowadays, technological advancements and modern chatting applications used for exchanging text messages have made information like emojis or stickers a reliable tool for identifying sender. Characteristics such as speech style or common phrases can also help determine the speaker.

It should be noted that, when analysing daily conversations, a large amount of information is required to identify a particular speaker from a typical phrase, or in other words, to identify their idiosyncrasies and style of speech. However, storytelling mediums, like movies, TV shows, books or video games, especially those set in the realm of fantasy, often accentuate these idiosyncrasies to increase immersion and make the story more believable. As such, using dialogues from these storytelling mediums provides a valuable case study, as they tend to be more data efficient. Another benefit of this approach is the absence of privacy concerns for fictional characters as opposed to transcripts of real-life conversations.

I now present my work which consists of performing two different sub-tasks of speaker identification: the first being the classification of utterances to a single speaker, focusing on character-specific traits, and the second being the mapping of utterances to a class of speakers, in turn detecting class-specific traits. For this, I used two datasets, one that is publicly available as part of the LIGHT (abbreviation for **L**earning in **I**nteractive **G**ames with **H**umans and **T**ext) research platform [86] and another one gathered from a commercial video game *Dragon Age: Origins* using publicly available tools to acquire the said data. Due to the difference in number of utterances available per speaker, I decided to use the dataset acquired from *Dragon Age: Origins* to identify individual speakers and the LIGHT dataset to identify the class of speakers. My approach has shown an accuracy of 90.27% when identifying a class of speaker and of 74.86% on an individual character identification level, outperforming the methods used as baseline.

3.2 Related Research in Speaker Identification

Most of the work in the field of speaker identification is centred around speech or signal processing [87], however, the work of Mairesse et al. [47] has shown that textual features can also reflect the personality of the speaker – suggesting that a text-dependent approach can be beneficial to speaker identification. To the best of my knowledge, the experimental works that solely focus on the textual features for the task of speaker identification are few, with works of Serban and Pineau [88],

Kundu et al. [89] and Ma et al. [90] being the most prominent ones.

Serban and Pineau [88] proposed using logistic regression and recurrent neural networks to detect changes in speakers using movie dialogue dataset. Their task differs from my research as it focuses on detecting turn changes (i.e. switch between interlocutors) rather than identifying actual speakers. Another work focusing on movie dialogues is that of Kundu et al. [89], which was the starting point for my experiments. They have suggested using various classification algorithms, such as Naive Bayes (NB), K-Nearest Neighbors (KNN) and Conditional Random Forest (CRF) on vectors formed out of certain stylistic features, such as the number of adverbs or adjectives per word in each utterance. They relied on these approaches in order to identify the speaker on a dataset acquired from the movie script database[1]. The work of Ma et al. [90] used their results as a baseline and conducted additional experiments with Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) on transcripts from the TV show *Friends*, reporting improved results with CNN. However, to the best of my knowledge, no research has been conducted on text-dependent speaker identification that utilises video game data.

3.3 Data

Due to the limited amount of publicly available corpora that focus on dialogues stemming from video games, I have focused my experiments on the two following data sources:

1. A dataset of textual dialogues extracted from the *Dragon Age: Origins*[2] video game developed by the *Bioware* studios.
2. Dialogue data extracted from a large-scale, crowd-sourced, text adventure game dataset, developed as part of the LIGHT research platform, coming from the work of Urbanek et al.[86].

While different in their origin, both datasets have many thematic similarities because they are centred around dialogues revolving in a fantasy world. For the sake of better outlining their individual strengths and weaknesses, I will describe each one within a separate subsection.

3.3.1 Dragon Age: Origins Dialogue Dataset

Dragon Age: Origins is an action role-playing video game developed by *Bioware* studios in 2009. It features a large amount of dialogue between the player and Non-Playable Characters (NPCs for short) with different personalities, thus making it suitable for speaker identification. The text data was obtained using two resources: *Dragon Age Wiki*[3] – a fan website that contains information related to

[1] http://www.imsdb.com/; last accessed on the 26th of May 2023

[2] https://www.ea.com/games/dragon-age/dragon-age-origins; last accessed on the 26th of May 2023

[3] https://dragonage.fandom.com/; last accessed on the 26th of May 2023

the video game series, and *Dragon Age Toolset*[4], a software released by developers of the game in order to allow modification of the in-game content.

The *Dragon Age Toolset* comes with a Microsoft SQL Server[5] relational database that contains all of the video game resources, including dialogue utterances in textual form. Utterances were extracted from the database, pre-processed and stored in a dialogue file. Each utterance has a speaker and a primary listener (the person to whom the utterance is spoken), and an identifier of the setting under which the utterance occurs (part of a quest, a random encounter, etc.). The missing information, that is, the speaker's name or scenario in which the utterance occurred, was manually filled in by referencing the *Dragon Age Wiki* website. I separated and labelled utterances spoken by the player and ten of the NPCs accompanying the player on their journey. These NPCs, due to their status as companions, have many utterances associated with them and thus warrant a separate class. On the other hand, due to them being few in number, utterances belonging to all the other NPCs have been grouped into a single category labelled as 'Others'. Figure 3.1 gives a graphical representation of the dialogue distribution within the dataset.

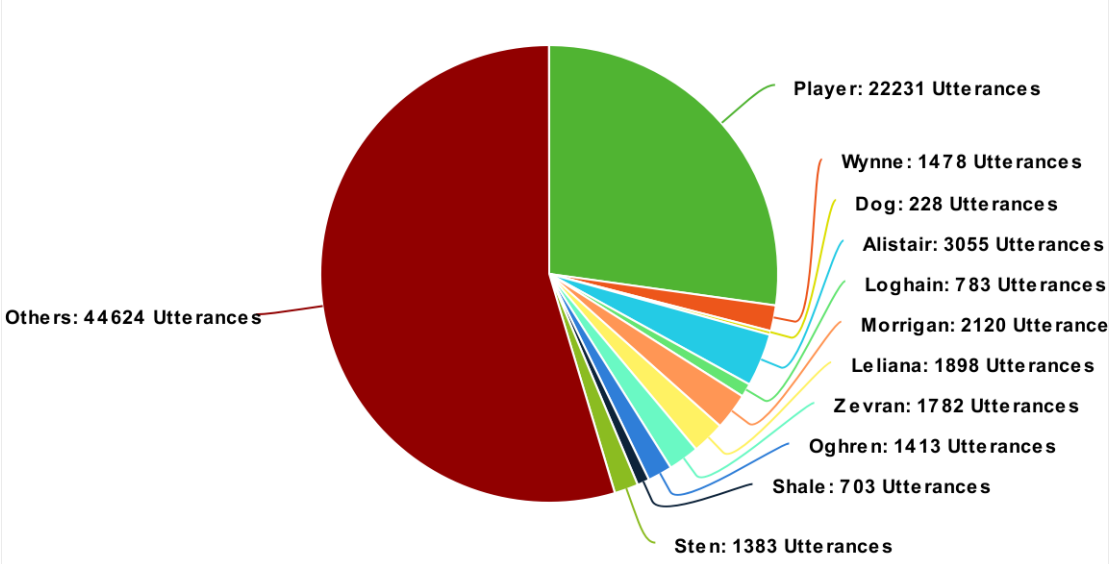

Figure 3.1: The number of utterances obtained from the *Dragon Age: Origins* game, separated by character. By far the largest category remains 'Others', while the 'Player' category amounts to roughly 25%.

[4] https://dragonage.fandom.com/wiki/Toolset; last accessed on the 26th of May 2023

[5] https://www.microsoft.com/en-au/sql-server/sql-server-downloads; last accessed on the 26th of May 2023

3.3.2 LIGHT Research Platform Dataset

LIGHT is a fantasy text adventure game research platform designed for studying grounded dialogue[6]. The dataset contains a large set of crowd-sourced interactions (about 11,000 episodes) consisting of actions and dialogue utterances. Data is publicly available through the ParlAI[7] platform. For this work I only considered the episodes which are part of the original LIGHT dataset. All of the utterances from episodes are extracted, pre-processed and stored in a dialogue file.

The LIGHT dataset contains many characters (940 characters with utterances), and unlike the other dataset I utilised, it lacks an evident main character as it does not follow a linear story. Additionally, most of the characters in the dataset have a character class assigned to them, with about 20% of them belonging to an undefined class. For unlabelled characters, I have manually assigned them one of the three existing classes: object, creature or person, based on their persona description which is also provided in the dataset.

For the above reasons, I turned to predicting character class from utterances rather than predicting a single character when using LIGHT dataset. Figure 3.2 shows utterance distribution per character class in the dataset.

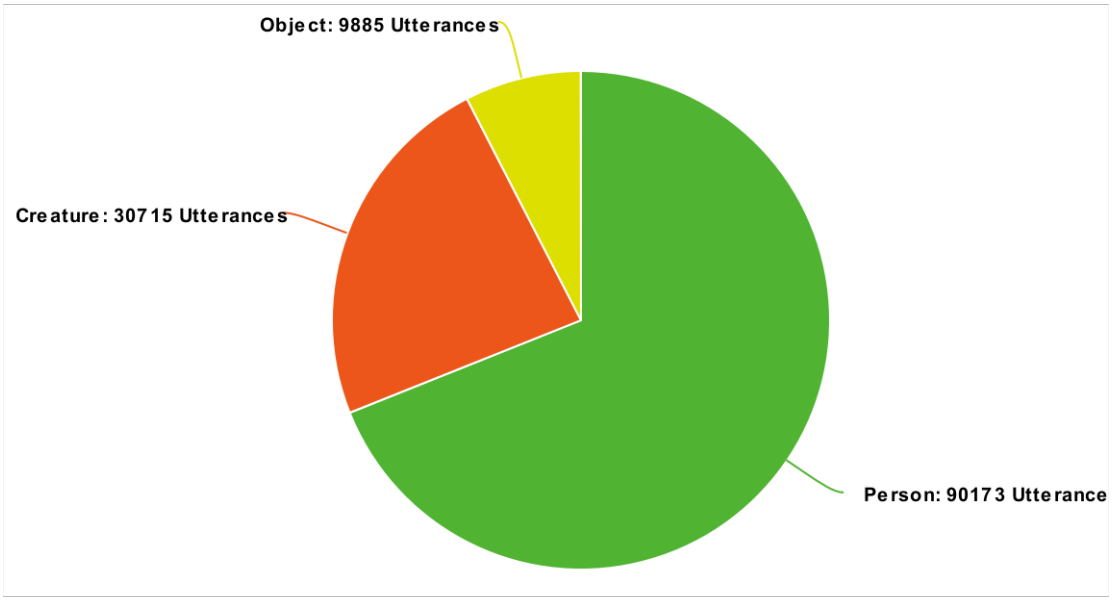

Figure 3.2: Number of utterances obtained from the LIGHT project dataset per character class. Much like the *Dragon Age: Origins* dataset, there is one class that amounts for majority of utterances.

[6]Grounded dialogue represents a type of conversation which is based on the shared understanding of the world between participants. As such it relies on concepts that participants can understand within a specific setting or context

[7]http://parl.ai/projects/light; last accessed on the 26th of May 2023

3.4 Methods

For methods, I have decided to focus on the four different approaches, three of which serve as a baseline for the final, proposed, approach. These baselines come from the works of Kundu et al. [89] and Ma et al. [90] previously described in the related works subsection. The proposed method is based on the **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) introduced by Devlin et al. [91]. For better overview, each of the mentioned approaches will be briefly introduced in this section.

3.4.1 K-Nearest Neighbors

In the work of Ma et al. [90] KNN was used as a baseline approach for speaker identification, while Kundu et al. [89] reported it to be the best-performing algorithm for the same task. Although other approaches proposed by Ma et al. outperform the KNN method, I have nonetheless decided to use it as a baseline due to its stability and simple implementation.

The algorithm has been implemented using the Facebook's Faiss[8] library for the Python programming language due to its speed and efficiency. The implementation followed the specifications given by Kundu et al., using the same stylistic features combined with a cosine similarity which serves as a distance function.

3.4.2 Convolutional Neural Network

Ma et al. have reported satisfactory results using a Convolutional Neural Network (**CNN**) model proposed by Kim [92] with minor modifications. I selected this approach as another baseline for my experiments. The model consists of a single convolution layer, followed by a global max-pooling layer and a fully-connected layer with a softmax function used to normalise the output. For the sake of simplicity, I have decided not to enforce the Euclidean (L^2) norm constraint, as the work of Zhang and Wallace [93] found that it has little effect on the final result.

Unlike the previous work, I have decided to add an embedding layer instead of using the pre-trained word embeddings due to a large number of colloquialisms and fantasy-related vocabulary present in the data, as well as the size of the data itself. This has made the training phase slower compared to previous work, as I also train the embeddings. In order to avoid possible overfitting, early stopping [94] is used as a form of regularisation.

3.4.3 Convolutional Neural Network with Utterance Concatenation

The final baseline method, much like the previously discussed one, is also based on the proposal of Ma et al.. It is an approach that involves using CNN while

[8] https://github.com/facebookresearch/faiss; last accessed on the 26th of May 2023

grouping a single character's utterances within a scene in chronological order. Ma et al. have reported that this approach achieves improved F1 and accuracy scores over the approach that does not rely on utterance concatenation on the speaker identification task.

In order to better replicate their experiments, I have grouped the utterances from a single episode of the LIGHT dataset, as it can be seen as a rough equivalent of a scene within a TV series. Similarly, when grouping the data from the *Dragon Age: Origins* video game, I considered utterances with the same *speaker* and *settingname* attribute. This is mainly because utterances with the same *settingname* attribute belong to the same dialogue tree[9], which can also be seen as equivalent to a TV series scene.

Ma et al. have also reported improvements for the method by restricting the prediction search space to only the set of speakers present within the observed TV series scene. However, as the majority of video game dialogues present in the data used for my experiments is limited to only two participants, I have decided to avoid replicating this step of their approach.

3.4.4 Bidirectional Encoder Representations from Transformers

BERT is a transformer-based model published in 2018 and has been applied ever since on various NLP tasks. It has achieved considerably better results than other approaches, especially on classification problems [95, 96]. The work of Chalkidis et al. [97] reported satisfactory results regarding multi-label text classification on documents, which inspired me to test it on a single utterance level.

For speaker and speaker class identification, I used an uncased BERT base model that utilises the Adam optimisation algorithm with weight decay. The experiments were run on a single utterance level and with concatenated texts for both datasets through 5 epochs. The dropout was set to 0.1 and learning rate to $1e-5$. Due to memory constraints, I used a batch size of 3 in my experiments.

3.5 Experiments

For my experiments, all the utterances extracted from both datasets are first cleaned and tokenised. With KNN-based and CNN-based methods, I have used a custom tokeniser that is a minor modification of that used in the work of Kim [92], while for the BERT model, I have used a BERT tokeniser. For the KNN method, 70% of the data was taken as a training set and 30% as an evaluation set. For other approaches, 70% of the data was used for training, while 15% was used for validation and the remaining 15% for evaluation set. Due to the nature of the data split, experiments were conducted over ten trials, taking the mean value as the final result.

I have experimentally established that for the KNN approach, results are

[9] https://en.wikipedia.org/wiki/Dialogue_tree; last accessed on the 26th of May 2023

best for k=13 on the *Dragon Age: Origins* dataset, while for the LIGHT dataset the results are best for k=15.

Figures 3.1 and 3.2 show that there is a severe class imbalance present in both of the datasets. This has led me to take an oversampling approach [98] for the KNN and CNN based methods in order to prevent classes with many examples from skewing the classifier output. However, since BERT has been shown to perform well even on an imbalanced datasets [99], for my experiments, I report the results that have been achieved without using any data augmentation techniques.

During the training phase of the CNN approaches, I used a batch size of 100 combined with the early stopping regularisation to determine the correct number of epochs needed for the training process while avoiding overfitting. For the BERT-based approach, I used five epochs. Table 3.1 displays the results, with the upper half showing results achieved on the *Dragon Age: Origins* dataset for the speaker identification task and the lower half reporting results on the LIGHT dataset for the speaker class identification task.

Model	**Precision**	**Recall**	**Accuracy**	**MF1**	**WF1**
***Dragon Age: Origins* dataset**					
KNN	3.604	10.329	20.045	4.613	8.065
CNN	43.656	36.253	39.662	36.785	37.302
CNN-Concatenation	36.332	27.958	45.672	24.850	34.776
BERT	46.431	52.757	**74.859**	49.060	**74.227**
BERT-Concatenation	**60.578**	**64.422**	70.212	**50.548**	69.302
LIGHT dataset					
KNN	38.579	33.743	76.328	30.117	67.002
CNN	**76.018**	52.675	79.912	49.853	72.483
CNN-Concatenation	53.168	49.339	83.292	50.681	62.124
BERT	48.828	63.673	83.417	50.883	82.588
BERT-Concatenation	70.669	**73.024**	**90.272**	**71.710**	**90.192**

Table 3.1: Performance per model. **MF1** stands for macro F1 score and **WF1** for F1 score weighted by the number of true labels per class.

3.6 Discussion

In the work of Kundu et al. [89] KNN was the best-performing algorithm on a movie script dataset, with an accuracy of 30.39%, while other metrics were not reported. The experiments of Ma et al. [90] had the most success using CNN in combination with utterance concatenation while also restricting the set of prediction labels to speakers present in a scene. They reported an accuracy of 46.48% and an F1 score of 44.19% when correctly predicting speakers from dialogue transcripts of the TV show *Friends*.

Using KNN and CNN-based approaches on the *Dragon Age: Origins* dataset proved to be less successful, with the CNN one yielding the best results out of all

the baseline methods. It should be noted that, surprisingly, concatenated utterances led to an increase in false positives and false negatives (lower recall and precision values) rather than improving over the simple CNN approach. After looking at the results, I believe that this is due to concatenated utterances being more challenging to discern from one another, as they share a common pool of words and topics.

The results of the experiments I conducted on the LIGHT dataset are better than those using the same methods on the *Dragon Age: Origins* dataset. This result is expected due to the smaller number of labels present for this sub-task, making it easier to connect an utterance to a character class than it would be to a single character. I want to note that even though using the CNN with concatenated utterances has led to higher accuracy, it also caused an increase in both false positives and false negatives (reduction in precision and recall). While this was surprising, since I tried to balance out the classes for the experiment, it is my understanding that synthetic examples led to an increase in generalisation, in turn resulting in a slightly worse-than-expected performance.

According to the results shown in Table 3.1, BERT outperforms baseline methods without any data augmentation approaches. However, BERT still produces a considerable amount of false positive and false negative predictions per class, as indicated by both the relatively low precision and recall. I believe this can be improved by using data augmentation techniques in the future alongside the BERT model. Figure 3.3 shows a confusion matrix for the characters present in the *Dragon Age: Origins* dataset. Looking at it, I can make a couple of interesting observations. For example, despite the utterances for the character 'Dog' originating from the class with the fewest utterances, the predictions seem to achieve an accuracy of almost 100%, which can be attributed to their specific nature (they are mostly onomatopeic). I believe this can indicate a good performance achieved by the model on a set of less generic utterances.

Another interesting observation from the confusion matrix is that the characters the model tends to confuse the most between seem to be those that express shared personality traits. For example, both the 'Leliana' and 'Alistair' are stereotypical good characters that tend to show disapproval for any of the actions by the 'Player' character that could be considered evil. I believe that, for this reason, the model has shown the second highest confusion in predictions between these two characters.

The highest confusion, however, seems to be between the 'Player' and 'Others' character labels. I believe this is primarily due to the wide variety of utterances available for selection in the dialogue options offered to the 'Player' character, all of which were considered in my experiments. Some tend to be mutually exclusive, meaning that if a player selects one of the utterances as a dialogue option, the other ones will not appear in the dialogue tree or as options further down the interaction with other NPC characters. This allows players to reflect different personality traits through their dialogue choices. However, since these utterances have not been separated in the dataset for the experiments I performed in my work, the 'Player' character most likely exhibits many different personality traits, some of which are even conflicting. This is similar to the 'Others' category, which groups a wide variety of different NPC characters, all of which share different backgrounds and exhibit different personality traits.

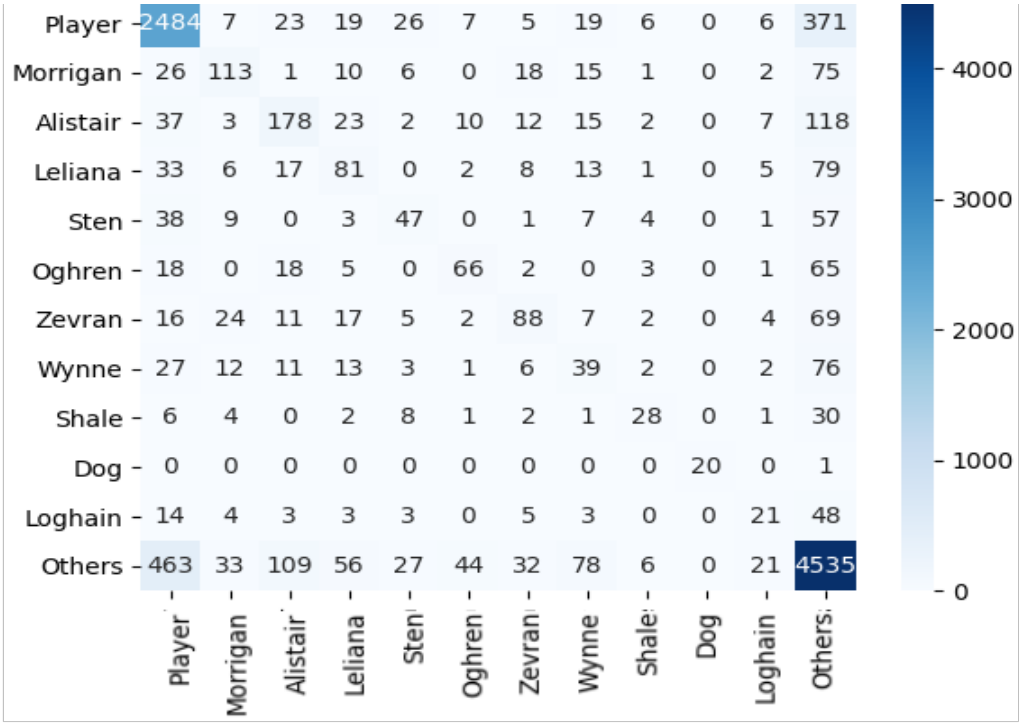

Figure 3.3: Confusion matrix for *Dragon Age: Origins* dataset on single utterance level.

3.6.1 Limitations of the Study

While this study offers valuable insight into the effectiveness of text for recognising interlocutors, it also has shown some limitations. BERT has been effectively used in many different experiments within the field of NLP. However, it is due to its black-box nature that it is difficult to claim that results reported in this study tend to be due to different personality traits or perhaps different linguistic structures used by characters. While the latter can be influenced by the former, there is no clear telling that the two are related without conducting a detailed analysis into the topic. This has been made difficult because there has yet to be any official information regarding these characters and their personality traits. Additionally, there are yet to be any studies in the field of psychology regarding this information within the video game domain. The only possible source of insight regarding different characters is offered by fan-operated websites (e.g. Personality-Database[10]) which tend to be limited in the amount or the quality of information offered.

Another study limitation can be seen in the sets of features considered for the experiments. While focusing on textual features alone has offered good insight into the validity of using text when recognising characters, results could be further improved by considering different features. For example, those that are more implicitly reflected in the text, such as contextual or personal information like age and pronouns. These features can perhaps be obtained from the *Dragon Age Wiki* website and used to enrich the dataset produced, making it more similar to the contribution

[10] https://www.personality-database.com/; last accessed on the 26th of May 2023

of Zhang et al. [100], known as the PERSONA-CHAT dataset. Experiments using these features could offer further insight by taking a similar approach to one proposed by Iosif and Mishra [101], who utilised speaker identification for multi-step analysis.

3.7 Conclusion

In this chapter, I described a new, transformer-based machine learning approach for the task of speaker identification. Even though most of the works in the field deal with audio and signal processing, the results reported using textual data have been promising. In order to accurately determine how well transformer-based approaches perform on the task, the results have been compared with different methods used in previous studies that focused on text-related information. These methods have been used as baseline approaches, with the experiments conducted by closely following the instructions provided by the authors of the baselines. The results confirm the findings of the previous works, namely the work of Ma et al. [90], that contextual information tends to be very important for the task of speaker identification, even in the domain of video game dialogues. When comparing the performance of different algorithms, an increase across most of the metrics has been observed when the utterance concatenation approach was employed to provide more context.

The corpus that has resulted from this research has proved to be a good starting point for the task. However, additional steps could be taken in order to improve the results. One of the examples of the improvements could be combining different commercial video games in a single corpus to diversify the pool of words used in dialogues. Although it should be noted that taking this approach could possibly introduce additional confusion for the model, as certain character tropes could be shared across multiple different works of fiction (e.g. stereotypical villains that use clichéd phrases).

In the case of the experiments conducted on the LIGHT dialogue dataset, the character classes considered could be further finely granulated to avoid imbalance between them. This could make it more difficult for the model to predict the correct class, as they would increase in number, but would, in turn, offer more insight into the similarities between characters with particular traits or similar backgrounds.

The final results suggest that it is possible to determine which utterance belongs to which character with significant accuracy when working with video game dialogue. Due to the entertaining nature of the domain, certain traits tend to be overly exaggerated, which has possibly assisted the transformer-based model in making more accurate predictions. It could be interesting to see the results of similar experiments on more text-driven games or even transcripts of real-life conversations.

Chapter 4

Analysing Personality in Textual Communication

The research described in the previous chapter has established that it is possible to connect an utterance to an interlocutor with a significant degree of certainty using textual data only. However, while it did answer the question of whether it is possible, it did not offer any insight into why it is possible. The confusion matrix presented in Figure 3.3 has indicated that the highest confusion rate for the transformer-based model has occurred between the characters that seem to share similar personality traits. Considering this, along with the fact that personality psychology assigns a pivotal role to the concept of personality when it comes to how people express themselves and are perceived by others [11, 12], it is necessary to research the relationship between text-based features and personality traits. In this chapter, I detail my investigation into this relationship using a dataset that includes texts from social media platform and personality measurements for several personality models previously introduced in Chapter 2.

The main contributions of this chapter can be summarised in the following way:

1. The study further substantiates the linguistic properties of the Big Five model while offering a novel insight into the possible relationship between the Myers-Briggs Type Indicator and several linguistic features.
2. The results reported help further understand the connection between the Big Five model and the Myers-Briggs Type Indicator.
3. To the best of my knowledge, this is the first study that compares these two personality models from a linguistic perspective.

The following peer-reviewed publication has served as the basis for this chapter:

- Radisavljević, Dušan, Bojan Batalo, Rafal Rzepka, and Kenji Araki. "Myers-Briggs Type Indicator and the Big Five Model-How Our Personality Affects Language Use." In 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1-6. IEEE, 2022. [72]

4.1 Introduction

Personality plays an essential role in one's life. It explains the consistency and individuality of one's behaviour and shapes their interactions. As such, its effect on how people communicate has been thoroughly studied. However, most of these studies are centred around the well-established Big Five personality model, resulting in the neglect of other methods used for assessing personality.

While the Big Five has been a popular option in academic circles, other personality models have seen application in different fields (e.g. MBTI in consulting [33]). Due to this, the problem arises when there is a need for psychological analysis of the data that originates from the fields that use different personality models, as most of the relevant research and its insight are focused solely on Big Five personality measures.

While the personality models measure personality using different sets of measures, there has been some research that was able to find a connection between some measurements from one model to another. A good example is the relationship between some measures in the MBTI model with the measures offered by the Big Five, which several prominent studies [10, 102, 103] have reported on. Namely, the Big Five's *Extroversion*, *Openness*, *Agreeableness* and *Conscientiousness* dimensions were found to be correlated to the *Introvert/Extrovert*, *Sensing/Intuition*, *Thinking/Feeling* and *Judging/Perceiving* types, respectively. The only measure offered by the Big Five model that has not shown any correlation with the existing MBTI types has been the *Neuroticism* dimension (4.1 shows a graphic depiction of this relationship). This relationship can be highly beneficial when seeking to utilise personality research in consulting, where MBTI has been a popular choice for reporting one's personality. On the other hand, with the research of Celli and Lepri [78] suggesting easier access to MBTI labels, the correlation between two sets of measures can also prove beneficial for any research involving the Big Five model. This can also help minimise concerns like the high cost of labelling and privacy issues, which have frequently been cited as problems in the personality computing research field [70].

In this chapter, I further investigate this relationship between the two models by analysing the language used on the social media platform Reddit[1]. In my experiments, I utilise Linguistic Inquiry and Word Count (LIWC) [104] software – a tool commonly used in language analysis. I apply LIWC to better understand how individuals labelled with different personalities express themselves and what the differences between the personality measures belonging to the two models are. To the best of my knowledge, this is the first in-depth analysis of the language usage of individuals labelled with MBTI types and the first direct comparison between the MBTI and Big Five models from a language use perspective.

[1] https://www.reddit.com/; last accessed on the 26th of May 2023

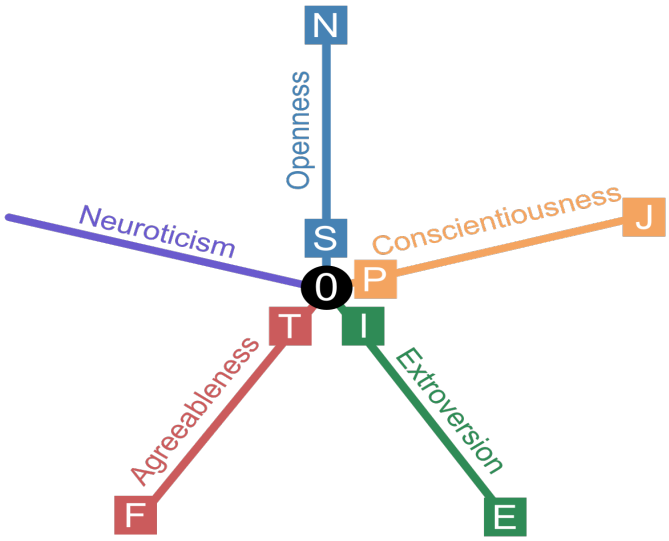

Figure 4.1: The relationship between the MBTI measurements (squares) and Big Five ones (lines) from the Big Five personality model perspective, with the *Neuroticism* trait showing no correlation to any of the MBTI types. The MBTI type closer to 0 represents a negative correlation with the Big Five measurement, on top of which it is found.

4.2 Related Research in Personality and Language Usage

When it comes to personality research, most of the available works focus on the Big Five model. The lack of popularity of the MBTI can be explained by various studies that have been critical of its shortcomings [10, 105, 106, 107], especially when compared to the Big Five model. However, with the popularity of the MBTI on social media platforms, some works have sought to utilise it in their studies, e.g., the works of Plank and Hovy [75] and Gjurković and Šnajder [70]. Despite this, the Big Five remains the more popular choice. Due to this difference in preference, as well as the Big Five's lexical background [29], most of the similar research focuses on analysing the relationship between linguistic features and the Big Five model.

While several different sets of features can be used to analyse the connection between language usage and personality [47], the most popular one has been LIWC. Since its inception in the mid-'90's, LIWC has been one of the most popular tools for language analysis. It has been applied to a wide variety of topics, such as research into political affiliation [108], depression [109] and sarcasm detection [110]. Apart from these research areas, many works relied on LIWC for personality analysis and prediction [47, 111, 112]. When it comes to the works that analyse the relationship between LIWC and the Big Five model, some of the most notable examples are the work of Pennebaker and King [48], Mehl et al. [113] and that of Holtgraves [114].

The work of Pennebaker and King [48] was a study organised into three parts.

During the third part of the study, they tried to establish a relationship between dimensions of language offered by LIWC and traditional measures of motivation, such as the "need" dimensions associated with the Thematic Apperception Test (TAT) [115], demographic information and the Big Five model measurements. On the other hand, Mehl et al. [113] used EAR dataset and Big Five personality framework to test the extent to which the behavioural manifestation of the participants and the implicit folk theories agree with their personality. Regarding language usage, they relied on several categories offered by LIWC. Finally, the work of Holtgraves [114] measured the correlation between several LIWC categories and the Big Five personality traits among participants using text messages. However, all of these studies differ from mine as they solely focus on the relationship between the Big Five model and LIWC categories.

4.3 Analysis and Results

Since the data coming from the Reddit social media platform is most similar in format to that of text messages, for my experimental setting I replicated the approach taken by Holtgraves in their work that previously examined the relationship between LIWC features and the features of the Big Five model [114]. Thus, I can compare my results with the ones they reported, with hopes of additionally uncovering the relationship between the LIWC features, the user's reported gender and personality traits.

I split the data into groups of users labelled either with MBTI personality traits or Big Five dimensions. This has resulted in $8,691$ users and $15,597,237$ comments labelled with MBTI types, while the group with Big Five dimensions amounted to $1,175$ users and $3,006,566$ comments in total. I followed this by grouping each user's comments, allowing for the LIWC analysis to be conducted on an individual level and on the level of the personality model.

4.3.1 Big Five Analysis

Out of the $1,175$ users labelled with Big Five dimensions in the PANDORA[2] dataset [30], I first extracted a group of 599 users that also reported their gender. This gave me a subset of 232 female and 367 male users. For comparison purposes, Table 4.1 reports the correlation with the Big Five's *Agreeableness* dimension, Table 4.2 for *Extroversion* and Table 4.3 for *Neuroticism*. I observed that most LIWC dimensions with significant correlations in the previous work also appear in these tables.

To better capture the nature of human language usage in texting, Holtgraves introduced specific categories by expanding the LIWC dictionary, such as slang, acronyms, and words that have their g's at the end dropped (e.g. goin'). I decided to skip this step, as the frequency of such terms in the PANDORA dataset has led me to conclude that there is no need for a separate category. Additionally, I selected two more LIWC dimensions I found interesting for analysis: the usage of *inclusive*

[2] **P**ersonality **AN**d **D**emographics **O**f **R**eddit **A**uthors abbrev.

and usage of *exclusive* language. Further discussion of the results is offered towards the end of this section.

Table 4.1: Correlation between the *Agreeableness* trait and LIWC categories

LIWC	**Agreeableness**		
Dimensions	Overall	Female	Male
Word count	0.02	0.051^{**}	0.004
Personal pronouns	0.091^{**}	0.109^{**}	0.029
1st person singular	0.054^{**}	0.052^{**}	0.005
Impersonal pronouns	0.077^{**}	0.093^{**}	0.067^{**}
Positive emotions	0.177^{***}	0.265^{***}	0.114^{**}
Negative emotions	-0.096^{**}	-0.023	-0.127^{**}
Anxiety	0.056^{**}	0.055^{**}	0.024
Anger	-0.193^{***}	-0.143^{**}	-0.202^{***}
Swearing	-0.123^{***}	-0.075^{**}	-0.131^{**}
Sex	-0.043^{**}	0.009	-0.116^{**}
Function Words	0.124^{***}	0.146^{**}	0.091^{**}
Prepositions	0.109^{***}	0.054^{**}	0.135^{***}
Death	-0.083^{**}	-0.108^{**}	-0.047^{**}
Health	-0.007	-0.027	-0.012
Inclusive	0.107^{***}	0.145^{**}	0.053^{**}
Exclusive	0.07^{**}	0.065^{**}	0.064^{**}

* when $p < 0.1$, ** when $p < 0.05$, *** when $p < 0.01$

The analysis has pointed to several LIWC features positively correlating with the *Extroversion* dimension. Namely, the results have suggested that people with a higher value for this Big Five dimension tend to use more words ($r = 0.066$), and personal pronouns when commenting ($r = 0.062$). This trend has been especially apparent in users labelled with the female gender ($r = 0.137$ and $r = 0.177$, respectively). These results agree with those previously reported by Holtgraves, as they have also observed a similar trend. Additionally, a positive correlation is observed with words that are often associated with *sex* ($r = 0.054$). This trend was also present in the results of previous research. However, while previous work has pointed to a negative correlation with words relating to *anxiety* and *anger*, my analysis indicated that users with higher *Extroversion* measures are more prone to using such words ($r = 0.043$ and $r = 0.039$, respectively).

For the *Neuroticism* dimension, I observed that people with higher values tended to avoid words associated with *anger* ($r = -0.067$) and used fewer *swear* words ($r = -0.057$), while also avoiding the usage of both *inclusive* ($r = -0.089$) and *exclusive* ($r = -0.082$) language. In addition, I noted that users with higher *Neuroticism* were less likely to use words tied to *positive emotions* ($r = -0.047$ for female and $r = -0.048$ for male gender). While the work of Holtgraves has reported the opposite to be the case, in their discussion chapter they stated that a negative correlation between the two dimensions was expected.

Finally, in this analysis, *Agreeableness* has shown a negative correlation with usage of *swear* words ($r = -0.123$) and words relating to *negative emotions* ($r = -0.096$). This was previously noted to be expected by Holtgraves in their paper. I further observed the *Agreeableness* dimension correlating with the usage of *imper-*

Table 4.2: Correlation between the *Extroversion* trait and the LIWC categories

LIWC	**Extroversion**		
Dimensions	Overall	Female	Male
Word count	0.066^{**}	0.137^{**}	0.019
Personal pronouns	0.062^{**}	0.177^{***}	-0.034
1st person singular	-0.002	0.083^{**}	-0.09^{**}
Impersonal pronouns	-0.102^{**}	-0.102^{**}	-0.104^{**}
Positive emotions	0.073^{**}	0.09^{**}	0.056^{**}
Negative emotions	-0.003	0.047^{**}	-0.024
Anxiety	0.043^{**}	0.163^{**}	-0.048^{**}
Anger	0.039^{**}	0.016	0.066^{**}
Swearing	0.019	0.007	0.034
Sex	0.054^{**}	0.041	0.054^{**}
Function Words	-0.014	0.015	-0.043^{**}
Prepositions	0.035^{**}	0.051^{**}	0.023
Death	-0.047^{**}	-0.121^{**}	0.027
Health	0.03^{**}	0.058^{**}	0.004
Inclusive	0.13^{***}	0.095^{**}	0.142^{***}
Exclusive	-0.122^{***}	-0.088^{**}	-0.148^{***}

* when $p < 0.1$, ** when $p < 0.05$, *** when $p < 0.01$

sonal pronouns ($r = 0.077$) and *positive emotions* ($r = 0.177$), while words relating to *death* were negatively correlated ($r = -0.083$). Additionally, users with high *Agreeableness* were more likely to use *inclusive* ($r = 0.107$) and *exclusive* ($r = 0.07$) language.

Previous work has not reported a significant correlation between *Conscientiousness* and *Openness* with any of the LIWC dimensions; however, my analysis has indicated that using *personal pronouns* seems negatively correlated with high *Openness*. Additionally, female-labelled Reddit users that scored highly on the *Conscientiousness* dimension are more likely to use *inclusive* language, while no significant correlation was observed in users that listed their gender as male.

4.3.2 MBTI Analysis

For the MBTI analysis, I focused on the subset of 2,695 users that reported their gender (1,184 female and 1,511 male). To directly compare the results to those achieved when using the Big Five personality traits, I selected the same set of LIWC categories used in the previous subsection. Tables 4.4, 4.4, 4.6 and 4.7 report the correlation scores for the *Introvert*, *iNtuitive*, *Feeling* and *Perceiving* types, respectively. In addition to the LIWC dimensions analysed in the previous subsection, I included reports on the correlations with *working*, *achievement* and *leisure* dimensions for the *Perceiving* type, as some authors have previously reported a correlation between them [116].

Due to MBTI types being dichotomies, I report correlations only for a single value (e.g. for *Introvert* not for *Extrovert*). The results for the other type value are obtainable by multiplying the correlation reported by -1.

Table 4.3: Correlation between the *Neuroticism* trait and the LIWC categories

LIWC	**Neuroticism**		
Dimensions	Overall	Female	Male
Word count	-0.009	-0.015	-0.001
Personal pronouns	-0.033^{**}	-0.039	-0.088^{**}
1st person singular	0.063^{**}	0.013	0.061^{**}
Impersonal pronouns	0.029^{**}	0.085^{**}	-0.003
Positive emotions	-0.005	0.047^{**}	-0.048^{**}
Negative emotions	0.011	0.031	0.005
Anxiety	0.095^{**}	0.067^{**}	0.09^{**}
Anger	-0.067^{**}	-0.057^{**}	-0.054^{**}
Swearing	-0.057^{**}	-0.042	-0.052^{**}
Sex	-0.002	0.026	-0.047^{**}
Function Words	-0.063^{**}	-0.026	-0.104^{**}
Prepositions	-0.073^{**}	-0.094^{**}	-0.067^{**}
Death	0.011	0.04	-0.001
Health	0.01	-0.026	0.019
Inclusive	-0.089^{**}	-0.11^{**}	-0.109^{**}
Exclusive	-0.082^{**}	-0.057^{**}	-0.106^{**}

* when $p < 0.1$, ** when $p < 0.05$, *** when $p < 0.01$

Following the results that were reported for Big Five traits, one might expect the *Introvert* type to show a negative correlation with LIWC dimensions such as *positive emotions*, *word count* and *personal pronouns*. However, no significant correlation was observed for these three dimensions. On the other hand, my analysis has shown that the language of users labelled with the *Introvert* type is negatively correlated with words that relate to topics like *sex* ($r = -0.041$), as well as with words that are associated with *anger* ($r = -0.058$) and *swear* words ($r = -0.073$). Additionally, users labelled with this type seem to use more *impersonal pronouns* ($r = 0.055$), while there has been no significant correlation in the usage of *personal pronouns*.

As the *iNtuitive* type was proven highly correlated with the *Openness* dimension [10], no significant correlation was expected with any linguistic features. However, my analysis has indicated a marginally significant negative correlation concerning the usage of *personal pronouns* ($r = -0.078$), similar to the results observed in the previous subsection. Furthermore, users labelled as *iNtuitive* tend to use words associated with *positive emotions* slightly more than chance ($r = 0.044$).

Users labelled with the *Feeling* type have shown a positive correlation with language rich in *personal pronouns* ($r = 0.214$) as well as *impersonal pronouns* ($r = 0.209$). Additionally, I have observed a negative correlation regarding the usage of *swear* words ($r = -0.114$). This is an expected result, as a similar correlation has been observed for the *Agreeableness* Big Five dimension in the previous subsection, with these two personality measures having been shown to correlate in the past [10]. Other dimensions showing a marginally significant correlation with this MBTI type are *inclusive* language ($r = 0.138$) and *function words* ($r = 0.125$).

While I have observed a negative correlation between *work* ($r = -0.043$)

Table 4.4: Correlation between MBTI's *Introvert* type and the LIWC categories

LIWC	**Introvert**		
Dimensions	Overall	Female	Male
Word count	0.0	−0.008	0.005
Personal pronouns	0.002	0.017	−0.016
1st person singular	0.037**	0.046**	0.031**
Impersonal pronouns	0.055***	0.071**	0.044**
Positive emotions	0.006	−0.001	0.009
Negative emotions	−0.038**	0.011	−0.058**
Anxiety	0.037**	0.036**	0.038**
Anger	−0.058***	−0.036**	−0.071***
Swearing	−0.073***	−0.062**	−0.079***
Sex	−0.041**	−0.049**	−0.047**
Function Words	0.037**	0.048**	0.032**
Prepositions	0.02**	0.009	0.027**
Death	0.031**	0.048**	0.026**
Health	0.033**	0.038**	0.028**
Inclusive	−0.01	0.002	−0.025**
Exclusive	0.026**	0.009	0.036**

* when $p < 0.1$, ** when $p < 0.05$, *** when $p < 0.01$

and *achievement* ($r = -0.056$) related language and the *Perceiving* type, (something that was also observed in another research [116]) my analysis has shown no significant correlation with the *leisure* dimension ($r = 0.001$). However, it is interesting to note that users labelled with the *Perceiving* type have shown a correlation with several LIWC dimensions, such as words relating to *death* ($r = 0.086$), *sex* ($r = 0.043$), *negative emotions* ($r = 0.068$), *swear* words ($r = 0.129$) and *anger* ($r = 0.078$). Additionally, a negative correlation has been observed for the usage of *inclusive* words ($r = -0.118$), words relating to *positive emotions* ($r = -0.056$), *function words* ($r = -0.058$) and *prepositions* ($r = -0.086$). This is interesting to note as the *Conscientiousness*, the Big Five dimension with which the *Perceiving* type has a negative correlation [10], has shown no prevailing pattern relating to language use.

4.4 Conclusion

The study outlined in this chapter showcases the relationship between linguistic features in the language used on social media platforms and their correlation with the personality measures offered by two distinct personality models – the Big Five and the MBTI model. In order to give an accurate comparison to prior work done solely on the Big Five model, I focused on the users that have reported their gender, offering additional insight from this standpoint. This study offers a novel insight into language usage and personality measured by these two models while comparing them.

However, this study is not without its limitations. The main issue is that it focuses solely on the subset of data involving users publicly reporting their gender. This is problematic, as the willingness to disclose this information could be influ-

Table 4.5: Correlation between MBTI's *iNtuitive* type and the LIWC categories

LIWC	**iNtuitive**		
Dimensions	Overall	Female	Male
Word count	-0.016^{**}	-0.044^{**}	-0.009
Personal pronouns	-0.078^{***}	-0.06^{**}	-0.028^{**}
1st person singular	-0.075^{***}	-0.054^{**}	-0.032^{**}
Impersonal pronouns	0.035^{**}	0.071^{**}	0.013
Positive emotions	0.044^{**}	0.075^{***}	0.048^{**}
Negative emotions	0.007	0.008	0.006
Anxiety	0.004	0.035^{**}	0.042^{**}
Anger	0.006	-0.015	-0.001
Swearing	-0.019^{**}	-0.03^{**}	-0.04^{**}
Sex	0.007	0.02^{**}	0.009
Function Words	-0.019^{**}	0.005	0.007
Prepositions	0.046^{**}	0.072^{**}	0.036^{**}
Death	0.025^{**}	0.013	-0.003
Health	0.014^{**}	0.043^{**}	0.039^{**}
Inclusive	-0.008	0.004	0.037^{**}
Exclusive	-0.035^{**}	-0.051^{**}	0.005

* when $p < 0.1$, ** when $p < 0.05$, *** when $p < 0.01$

enced by several factors, including personality. As such, there is a potential for a certain amount of bias being introduced to the study, as the connection between the propensity to report gender and personality has not been examined. However, this approach was necessary in order to achieve an accurate comparison with the previous study conducted in the research field.

The findings of the analysis described in this chapter suggest that it is possible to use the linguistic connection between the MBTI and LIWC categories and its connection to the Big Five model to compare and possibly even translate from one set of measurements into another.

Table 4.6: Correlation between MBTI's *Feeling* type and the LIWC categories

LIWC	**Feeling**		
Dimensions	Overall	Female	Male
Word count	-0.11^{***}	-0.08^{***}	-0.119^{***}
Personal pronouns	0.214^{***}	0.196^{***}	0.187^{***}
1st person singular	0.209^{***}	0.171^{***}	0.197^{***}
Impersonal pronouns	0.112^{***}	0.091^{***}	0.122^{***}
Positive emotions	0.267^{***}	0.22^{***}	0.282^{***}
Negative emotions	-0.017^{**}	0.001	-0.023^{**}
Anxiety	0.193^{***}	0.145^{***}	0.192^{***}
Anger	-0.067^{***}	-0.086^{***}	-0.06^{**}
Swearing	-0.114^{***}	-0.086^{***}	-0.109^{***}
Sex	-0.005	0.069^{**}	-0.023^{**}
Function Words	0.125^{***}	0.079^{***}	0.117^{***}
Prepositions	0.03^{**}	-0.025^{**}	0.057^{**}
Death	-0.072^{***}	-0.039^{**}	-0.056^{**}
Health	0.03^{**}	0.013	-0.012
Inclusive	0.138^{***}	0.112^{***}	0.114^{***}
Exclusive	0.03^{**}	-0.026^{**}	0.045^{**}

* when $p < 0.1$, ** when $p < 0.05$, *** when $p < 0.01$

Table 4.7: Correlation between MBTI's *Perceiving* type and the LIWC categories

LIWC	**Perceiving**		
Dimensions	Overall	Female	Male
Word count	0.017^{**}	0.023^{**}	0.002
Personal pronouns	-0.068^{***}	-0.078^{***}	0.018^{**}
1st person singular	-0.048^{**}	-0.028^{**}	0.014
Impersonal pronouns	0.007	0.022^{**}	0.003
Positive emotions	-0.056^{***}	-0.06^{**}	-0.019^{**}
Negative emotions	0.068^{***}	0.087^{***}	0.066^{**}
Anxiety	-0.06^{***}	-0.034^{**}	-0.016
Anger	0.078^{***}	0.123^{***}	0.066^{**}
Swearing	0.129^{***}	0.105^{***}	0.12^{***}
Sex	0.043^{**}	0.068^{**}	0.05^{**}
Function Words	-0.058^{***}	-0.04^{**}	-0.028^{**}
Prepositions	-0.086^{***}	-0.094^{***}	-0.074^{***}
Death	0.086^{***}	0.113^{***}	0.033^{**}
Health	-0.053^{***}	-0.011	-0.038^{**}
Inclusive	-0.118^{***}	-0.131^{***}	-0.059^{**}
Exclusive	0.047^{**}	0.097^{***}	0.04^{**}
Work	-0.043^{**}	-0.028^{**}	-0.067^{***}
Achievement	-0.056^{***}	-0.118^{***}	-0.057^{**}
Leisure	0.001	0.002	-0.03^{**}

* when $p < 0.1$, ** when $p < 0.05$, *** when $p < 0.01$

Chapter 5

From MBTI Types to Big Five Traits

In the previous chapter, the results reported found a correlation between the MBTI measurements and several LIWC categories. This has indicated a possibility of leveraging this relationship and its connection to the similar relationship between the Big Five model and LIWC. Such a relationship could be beneficial to conversion from the more easily obtainable MBTI labels [78] into the more scientifically supported Big Five ones. However, due to the previously described analysis being conducted only on a limited set of people who have previously reported their gender, there is a need to investigate the relationship between these personality models further.

In this chapter, I seek to bridge the gap between the MBTI, Big Five and another personality model known as the Enneagram of Personality, to increase the number of resources for the Big Five model. I further explore the relationship which was reported between the MBTI types and certain Big Five traits, as well as test for the presence of a similar relationship between Enneagram and Big Five measures. As a result of these endeavours, the main contributions of this chapter can be summarised in the following way:

1. The series of detailed experiments that I conduct provides insight into the effectiveness of different features and regression algorithms for the task of personality prediction. Additionally, the choice of algorithms allows for greater interpretability of the results while maintaining a simplistic approach.

2. The proposed simple framework based on the psycholinguistic features that leverages the relationship between different personality models has led to an increase 0.033 points, or 13.2% in correlation strength, on the Pearson r correlation coefficient between predicted values and the gold-standard labels when compared to the baseline approach on a dimension-to-dimension level.

3. The use of psycholinguistic features helps further explore the relationship between language and how type-measured personality shapes its use in online spaces. While the relationship between the Big Five and language use has been thoroughly studied due to the lexical background of the Big Five model [29], similar studies for the type-based models are limited to the best of my knowledge.

The following peer-reviewed publication has served as the basis for this chap-

ter:

- Radisavljević, Dušan, Rafal Rzepka, and Kenji Araki. "Personality Types and Traits – Examining and Leveraging the Relationship between Different Personality Models for Mutual Prediction." Applied Sciences 13, no. 7 (2023): 4506. [117]

5.1 Introduction

The evaluation tools described in Section 2.3, also known as the personality models, often use different dimensions to measure an individual's personality. These dimensions, as well as their number, tend to vary on a model-to-model basis. However, it is possible to separate personality models into two larger groups based on the kind of value that they assign along the dimensions used. These groups are namely:

- *Trait-based personality models*, which use traits or, in other words, assign a continuous value along the dimension.
- *Type-based personality models*, which rely on types to describe a personality. Types can also be viewed as categories or discrete values selected from a dimension's domain.

When considering the personality models that have been previously introduced, the Big Five model would fit in the trait-based group. On the other hand, the Enneagram and the MBTI should be considered type-based personality models.

The process of assessing one's personality through one of these models has historically consisted of a psychological expert administering a test during an interview, with the test being dictated by the choice of the personality model. Though reliable and effective, this approach requires expert knowledge and is usually time-consuming. However, the combination of new machine-learning techniques and an increase in online communication has led to interest from scholars in the possibility of automating the tasks of prediction, interpretation and generation of dimensions that personality models use [118].

Coupled with advancements in new computational algorithms and the development of modern technologies, this has led to the formation of the *personality computing* research field. Another trend that can be considered a factor in the rapid development of this field is the increase in popularity of social media services, as this has encouraged people to share their interests publicly in online spaces [5].

However, while the amount of personality-related information increases daily, the need for more relevant personality-labelled data remains one of the most stated issues in the field [30]. This paradoxical phenomenon can be explained by the difference in personality model preference between academia and the non-psychological population that is more prevalent on social media. More specifically, while the Big Five model has seen extensive use in personality research, the MBTI has been the more popular choice for describing personality in online spaces.

The trait-based approach of the Big Five model, as well as its empirical support, cross-cultural applicability and reliability [119, 120, 121] have made it a popular choice in scientific circles, with a majority of research on personality computing centred around it. Several studies have additionally contributed to this preference by confirming its validity [122, 123], while the MBTI has often been criticised for lacking this evidence [10, 105, 106, 107]. If one considers the data that is publicly available through the Google Books n-gram Viewer API[1], they can note that although both the Big Five and MBTI personality models have seen an increase in popularity within the last couple of decades, the "Big Five" n-gram appears significantly more frequently when it comes to book titles (shown in Figure 5.1).

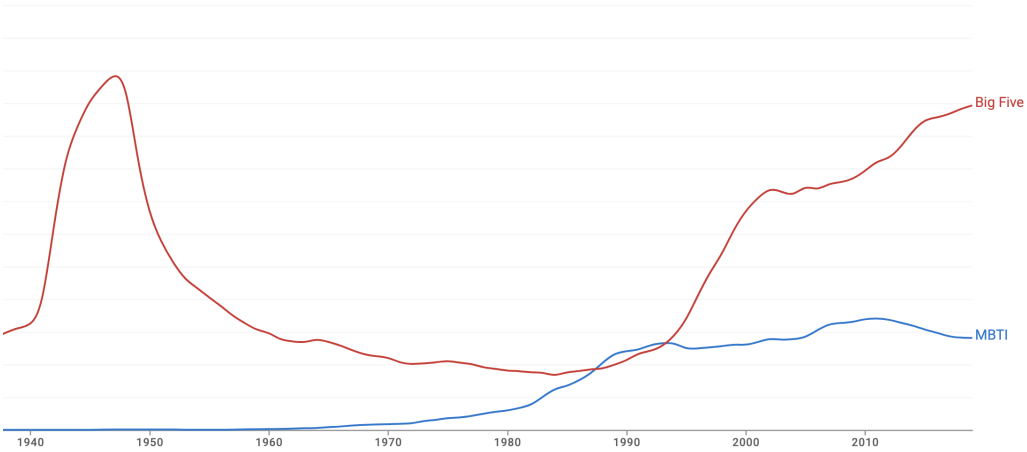

Figure 5.1: The frequency of personality models in book titles, showcasing interest in scientific circles for different personality models. The graph originates from the Google Books n-gram Viewer API. I note that the Big Five model is usually referred to as simply the "Big Five" in book titles.

On the other hand, if one uses the frequency with which people search for a particular personality model using the Google[2] search engine as an indicator of interest, it can be noted that the preference seems to be opposite from that in academia. When observing the data available through the Google Trends API[3], I noticed that the MBTI drew much more interest than the Big Five, especially in the last several years (Figure 5.2). This popularity can be attributed to how MBTI assigns personality – using a four-letter acronym. As such, it is easier to interpret and report for the non-psychological population, in turn causing it to be more prevalent on social media platforms and, thus, attract greater attention.

Despite the differences between the two previously mentioned personality models, several studies [10, 102, 103] have pointed towards a statistically significant correlation between the MBTI types and certain traits belonging to the Big Five model. This relationship between the two raises the question of whether it can somehow be leveraged to overcome each of their shortcomings. In the study described in this chapter, I aim to bridge the gap between these two personality models and, as a result, provide a significant increase in resources for the more scientifically accred-

[1] https://books.google.com/ngrams; last accessed on the 26th of May 2023
[2] https://www.google.com/; last accessed on the 26th of May 2023
[3] https://trends.google.com/home; last accessed on the 26th of May 2023

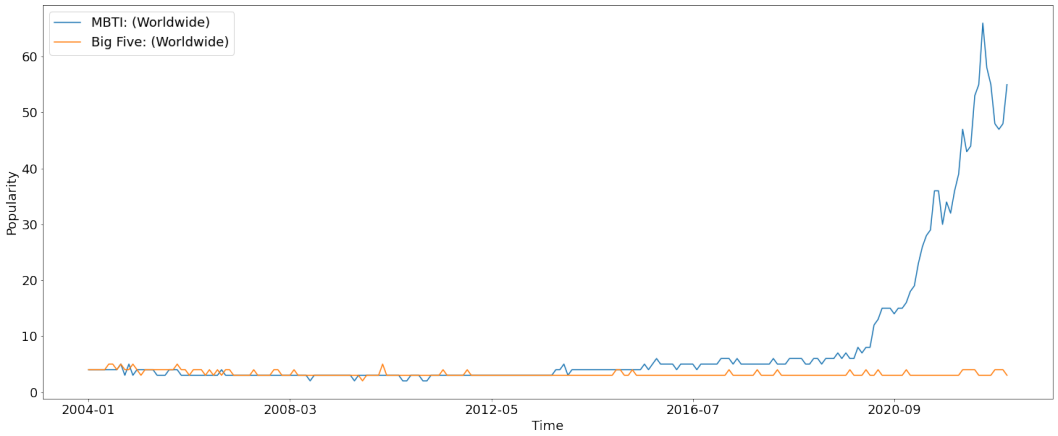

Figure 5.2: The frequency of personality models appearing as Google search terms, showcasing the interests of the general public. Data source: Google Trends API.

ited Big Five model, as well as bring scientific credibility to the often overlooked MBTI model.

The experiments that are described in this chapter further examine the relationship between the MBTI and the Big Five personality models. In addition, I look into the possible existence of a relationship between the Big Five and the Enneagram model, as well as the nature thereof. I conduct detailed experiments involving different sets of features and employ various regression algorithms, providing insight into their effectiveness on the task of personality prediction. Throughout this chapter, I adopt a comprehensive reporting style to assure replicability and better comparability with previous works, as the lack of comparison between studies has been a prevalent issue in personality computing [9]. The main goal of my experiments is to highlight the relationship between different personality models and further our understanding of personality [124].

5.2 Related Research in Personality Computing

A very influential work in personality computing is that of Mairesse et al. [47]. Their work used the EAR [113] and Essays [48] datasets to test the effectiveness of different features using classification, regression and ranking models. When using a smaller data set, they reported that simpler algorithms, such as Naive Bayes and regression trees, offered better performance; however, ranking models achieved better scores for a larger dataset. While their approach is closely related to my endeavours in analysing the effectiveness of different features, they focused on a single personality model. On the other hand, my work described in this chapter analyses the effectiveness of different features and how well they capture the relationship between multiple personality models rather than focusing only on the Big Five model and its relationship to the language usage.

An interesting study comparing multiple personality models is that of Celli and Lepri [78], who compared the effectiveness of predicting labels for the MBTI and the Big Five model on a dataset originating from the social media platform

Twitter[4]. They treated the problem as nine separate binary classification tasks, five of which were for classifying Big Five traits and four for the MBTI types. For the model architecture, they used a combination of n-gram, LIWC dimensions and metadata features for Support Vector Machine and another meta-classifier based on the work of Thornton et al. [125]. While their approach offered novel insight into the difference in effectiveness of the automatic personality recognition task for two different models, the actual relationship between models was left unexplored.

Several notable contributions have been made to the automatic personality recognition task using deep learning methods. Sun et al. [126] used a concatenation of bidirectional LSTMs and a convolutional neural network in order to predict personality traits from two Big Five personality datasets – the Essays dataset [48] and another one coming from the YouTube platform [81]. Kazameini et al. [127] also used the Essays dataset in their experiments; however, they adopted a multi-step approach that used a combination of Mairesse features [47] and BERT token representations [128] for prediction of the Big Five personality using a Support Vector Machine algorithm.

On the other hand, Kerz et al. [129] used a two-step approach that relied on BERT and BLSTM to predict Big Five personality traits from the Essays dataset and MBTI types from the MBTI Kaggle dataset [130]. While many more studies have used deep learning methods and showed promising results [8, 131, 132], most of them tend to focus on a single personality model – most commonly the Big Five.

In addition, the vast majority of studies that utilise deep learning methods treat the problem of Big Five trait prediction as a classification problem rather than a regression one. The only exception is the previously mentioned work of Kerz et al. [129], which focused on multiple personality models. However, due to Big Five traits and MBTI types labels originating from two different data sources, it is complicated to obtain more profound insight into the relationship between these two models.

My approach is heavily inspired by the work of Gjurković et al. [30], who introduced a dataset containing labels for the MBTI, Big Five and the Enneagram model stemming from a single social media platform. They were the first to explore the possibility of using labels from one personality model as features to increase the scores of the automatic personality recognition task for another. Their approach relied on a combination of MBTI and Enneagram predictions and a set of n-gram features to predict Big Five traits. In my work, I seek to extend their case study by taking a more detailed approach, similar to Mairesse et al. [47].

As with the works mentioned in this section, I focus on automatic personality recognition. However, to do so, I seek to leverage the relationship between multiple personality models and the way in which it is reflected through different features. The approach I take is detailed in nature for two reasons. The first is to assure comparability with approaches I used as a baseline, avoiding some common issues in the personality computing research field [7, 9]. The second reason is that, due to the complex nature of personality, a gradual introduction and experimentation with different features is the best way to accurately single out the effects they have on the task of personality recognition [47].

[4] https://twitter.com/; last accessed on the 26th of May 2023

5.3 Methods

When looking at tasks that use regression approaches, several methods can be utilised to improve results over the baseline. Some of the more common examples involve different regularisation methods, or data manipulation in the form of data augmentation and even data cleanup.

However, as the primary focus of the work described in this chapter is to improve our understanding of personality and how it is reflected across different personality models, I decided to use two methods that linearly approach the problem – (1) feature selection and (2) model selection. The exact design choices behind these methods will be further discussed in Subsections 5.3.2 and 5.3.3, respectively. To better contextualise the choices made, I start this section by briefly introducing the dataset used, as well as the approach that serves as a baseline for my experiments.

5.3.1 PANDORA Dataset and the Baseline Approach

While the dataset I use for my experiments has been previously briefly mentioned in Chapter 4, I would like to introduce it in greater detail within this subsection.

The PANDORA dataset stems from the social media platform Reddit, and is the contribution of Gjurković et al. [30]. As such, it represents a direct extension of their previous work that introduced the MBTI9k [70], another dataset containing MBTI labels. This dataset has been found to be suitable for my experiments due to the following reasons:

1. Most of the work done in personality computing that focuses on using data for different personality models utilises datasets from separate sources, with some examples being the works of Mehta et al. [133] and Kerz et al. [129]. On the other hand, in the work of other authors, while the data originates from the same source, it contains no overlap between users labelled with different personality models (e.g., the work of Celli and Lepri [78]). To the best of my knowledge, the PANDORA dataset is the **only dataset containing personality-relevant information for multiple personality models, with an overlap between user groups labelled with each of the models**.

2. The topical diversity of Reddit opens up the possibility of looking into the effects of interests and hobbies on personality prediction. Reddit is divided into a series of different "subreddits" – or smaller message boards. These message boards are often centred around a single topic or interest that individuals participating tend to share. **Information on these topical interests could be leveraged to improve the results of the personality prediction task**.

The PANDORA dataset consists of 10,288 users with labels for either the MBTI, Big Five or Enneagram personality models. In some cases, users have labels for multiple different personality models, causing an overlap between the labelled

groups. Additionally, some users have also stated their demographic information, such as gender or age. The dataset further includes 17,640,062 user comments written between January 2015 and May 2019. Table 5.1 gives an overview of the dataset, providing insight into the exact number of users and comments that contain labels for either MBTI, Big Five or Enneagram personality models, as well as information on the overlap between these groups.

Table 5.1: The number of users and comments labelled with each personality model in the PANDORA dataset. The data in this table was adapted from the work of Gjurković et al. [30] CC-BY-NC.

Personality Model	**Number of Users**	**Number of Comments**
Big Five	1,608	3,006,566
Enneagram	794	1,458,816
MBTI	9,084	15,597,237
Big Five and Enneagram	64	235,883
Big Five and MBTI	393	1,086,324
Enneagram and MBTI	793	1,457,625
All three models	63	234,692
Total	10,288	17,640,062

Observing how the data is distributed, I note that none of the personality traits follows a normal distribution (Figure 5.3). While this is not particularly unusual when analysing data, it should be noted that most previous works reported a tendency towards normal distribution for the Big Five traits (e.g., the works of Mairesse et al. [47] or the work of Uysal and Pohlmeier [134]). In the PANDORA dataset, however, most personality traits follow a skewed distribution, with the only exception being the *Neuroticism* trait, for which the labels seem to follow a bimodal distribution.

This phenomenon can be attributed to several reasons, such as selection bias [70] or the propensity towards openly stating personality traits being dictated by certain personality traits, e.g., high *Openness*. An additional possibility is that some subreddits and topics or interests tend to be more prevalent in the dataset; thus, the number of individuals with particular personality traits associated with such interests and topics tends to be higher. To test this possibility, I experimented with the effect that subreddit participation has on personality predictions, and the details of this study are described in Subsection 5.3.2.

While the data distribution of personality traits present in the dataset seems unusual (as seen in Figure 5.3), it should be noted that the correlations between the Big Five traits and the MBTI types reported in the data are largely in agreement with research that has previously examined the relationship between these two sets of dimensions [10, 103]. The only exception would be the *Openness* personality trait that, in the case of PANDORA, shows an unusually low correlation with the *S-N* type, despite previous works in the field reporting an agreement between these two dimensions that is higher than chance (Figure 5.4).

The approach used as a baseline in my work was described by the authors of PANDORA as a domain-adaptation task [135] of transferring the MBTI and

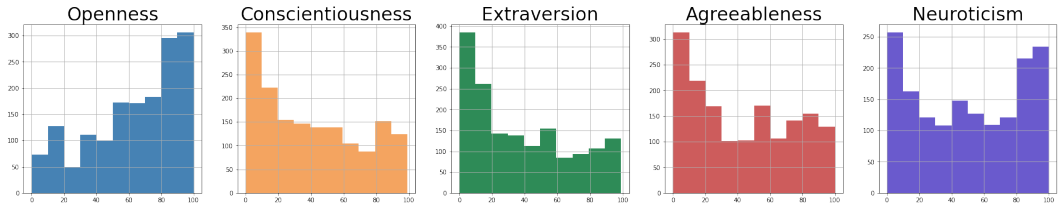

Figure 5.3: Probability distribution for each of the Big Five personality trait labels in the PANDORA dataset. While the previous research has suggested Big Five traits usually follow a normal distribution, personality traits of the PANDORA dataset [30] seem to follow a more skewed one.

Enneagram labels into the more scientifically relevant Big Five ones [30]. To this end, they started by training four logistic regression models [136] – one for each of the MBTI types and an additional one for the types present in the Enneagram. In order to train these models, they utilised a subset of MBTI/Enneagram labelled users only – in other words, a subset that had no overlap with users labelled with the Big Five personality traits. The set of users with labels for both the MBTI/Enneagram models and the Big Five were later used as part of a validation set.

The labels obtained from the five regression models were then used to predict MBTI/Enneagram values for the set of users without assigned labels for these type-based models. The predictions were then either used independently or combined with other features (e.g., gender, POS tags, stylistic features and named entities in text) and n-grams into a single feature set for the purpose of predicting the Big Five personality traits. The experiments were conducted using two different algorithms: (1) a linear regression model with an L^2 regularisation norm [137] (also known as Ridge regression) and (2) a trained neural network that utilised BERT [128] for text encoding. The results indicated better performance of the linear regression, with on average 0.15 higher results than those for the deep learning model.

5.3.2 Feature Selection Approach

The first step of my experiments focused on finding the optimal set of features for the task of automatic personality recognition. These features were then combined with the predictions of the type-based personality labels to leverage a relationship between them and the Big Five model. I theorise that the following three feature sources can benefit the results and lead to possible improvements over the baseline approach:

1. *Class predictions for the Big Five personality traits* – a set of features obtained from predicting Big Five labels as classes rather than values (e.g., "High Extroversion" instead of 74% *Extroversion*). The classes are constructed by applying a technique known as *binning*. These predictions were then combined with other features to predict the Big Five personality traits.

2. *Language-based features originating from Linguistic Inquiry and Word Count (LIWC)* – a set of psycholinguistic features produced as a result of statistical analysis conducted by the LIWC tool.

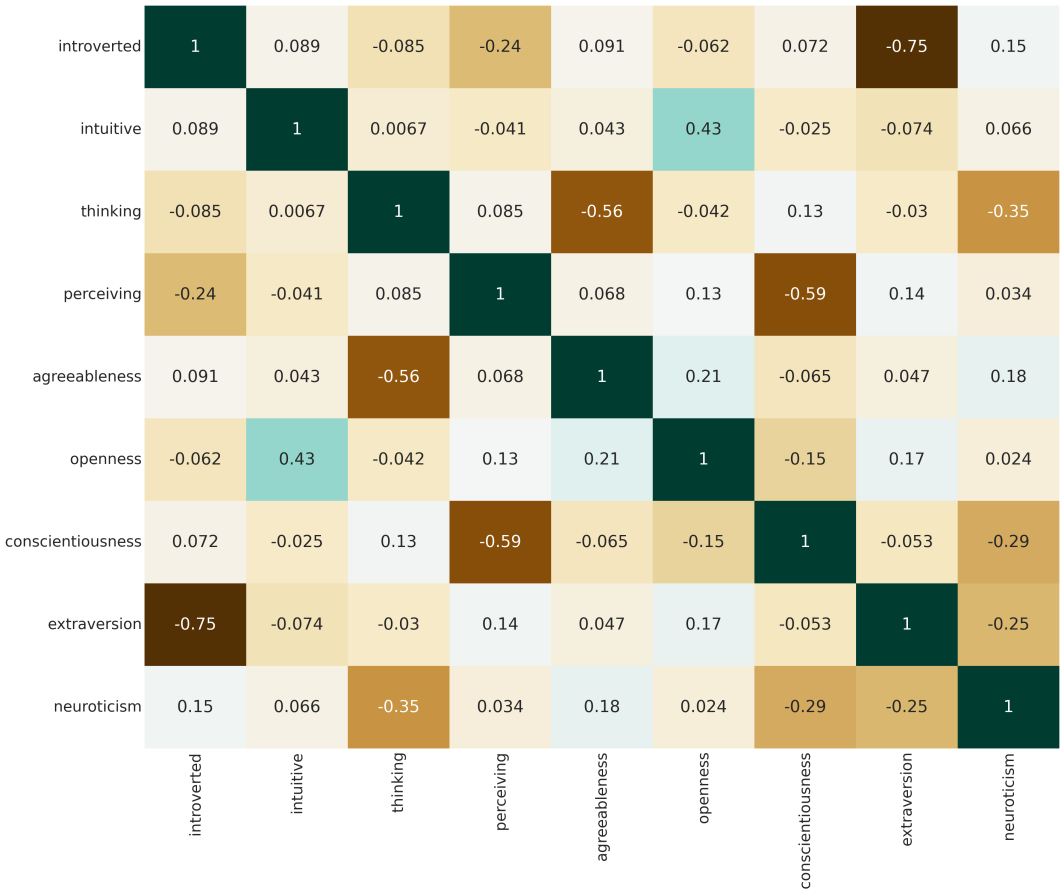

Figure 5.4: Heatmap indicating linear correlation between labels for the Big Five traits and MBTI types present in the PANDORA dataset [30].

3. *Information about user participation and engagement on the social media platform Reddit* – a feature set constructed from the frequency with which users post on the dataset's most and least popular message boards.

Big Five Classification Predictions as Features

My first hypothesis is based on the idea that the difference in the domain between personality models can ultimately impede the prediction results. While Enneagram types are represented using a whole number on a scale from 1 to 9, the MBTI types can be described using a binary value of either 0 or 1 in order to represent each of the four dichotomies aptly. On the other hand, Big Five personality traits are labelled using a positive number within the range of 0 to 100, with some labels even being represented as a single precision decimal number. It is possible to minimise these differences by introducing an additional step to the prediction process, which would treat the task as a classification problem rather than a regression one.

To convert the Big Five labels from continuous values into discrete ones, I have applied the binning method described by Segalin et al. [62], with slight modifications. In their work, the authors used two different techniques in order to separate Big Five traits into binary classes – (1) utilising the mean value of the particular

personality trait and (2) using the first and third quartiles of the distribution as class delimiters, discarding all values between them.

As the data distributions of the PANDORA dataset and the data used by Segalin et al. [62] differ, I introduced slight adjustments to the approaches they used.

Regaring technique (1) – instead of relying on the mean value, I used the median point of the personality trait distribution as a separator between classes. The reasoning behind this is that the median values tend to be more resilient to skewed data distributions, thus, making it a better fit for the PANDORA dataset (Figure 5.3). For technique (2), I decided against discarding any non-extreme value and instead binned the Big Five traits into three classes rather than two (high, low and middle). In doing so, I prevented any loss of information since the Big Five personality labels present in the PANDORA dataset are relatively smaller in size when compared to the number of MBTI ones (Table 5.1).

Despite the recent success of different deep learning approaches in predicting Big Five traits as classes [138, 127, 139], I decided to use the same regression algorithm as in the case of predicting MBTI and Enneagram types to allow for better comparability. The features used for this task include n-grams and MBTI/Enneagram predictions, as described in the work of Gjurković et al. [30]. These predictions were later used for the regression model, with Figure 5.5 illustrating the steps taken in predicting the continuous values for the Big Five personality traits.

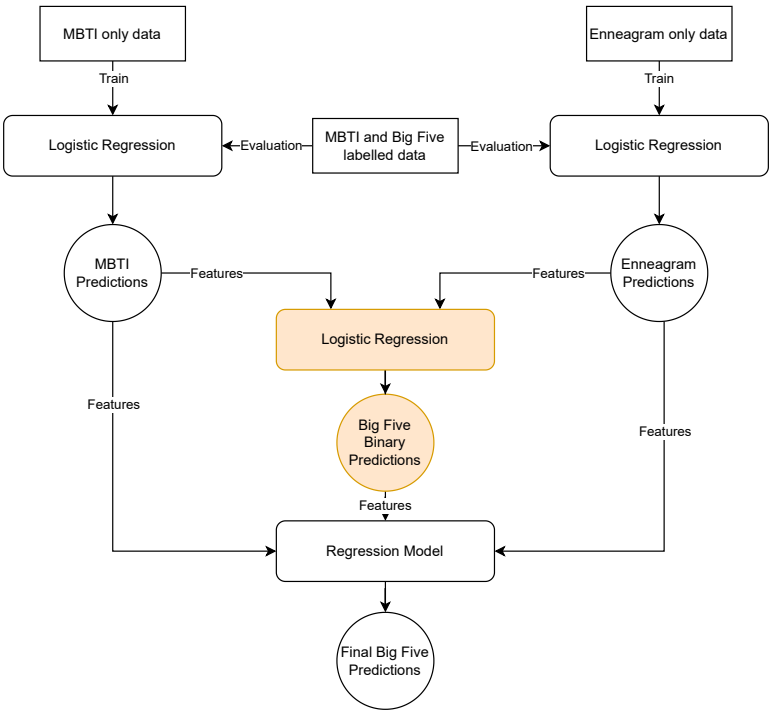

Figure 5.5: Illustration of the model stack after introducing the Big Five classification predictions. Orange highlights the added logistic regression model that predicts Big Five traits as one of two or three classes, depending on whether the median or quartile values are used as separators between classes. The newly added model acts as a weak learner in the model stack.

Language-Based Features

Boyd and Pennebaker [21] stated language to be one of the most important indicators of personality. As such, many linguistic and psycholinguistic features have seen extensive use in personality computing [140]. These language-based features often tend to be researched from the aspect of their relationship to the Big Five model (with some examples being the works of Mairesse et al. [47] and Holtgraves [114]). However, these features have rarely been used to connect the relationship between multiple personality models, with only few works attempting to do so [129, 72]. This is because the Big Five model directly results from a statistical analysis of the English lexicon [29], whereas the MBTI and Enneagram do not share a similar lexical background.

To further examine the relationship between these personality models and language, I rely on the Linguistic Inquiry and Word Count, also known as LIWC [104], which has been a popular tool for analysing how language is used. LIWC utilises over 100 internal dictionaries that test for the presence of various linguistic features, capturing the social and psychological states people express through language (a complete list of dimensions and their overview can be found at `https://www.liwc.app/help/psychometrics-manuals` (last accessed on the 13th of May 2023).

Each of these dictionaries consists of words, word stems, emoticons and other text features that help to identify the psychological category of interest from the textual data. For example, the "affiliation" dictionary contains about 350 entries, among which are words such as "community", "together" and other verbal constructs indicating a person's desire to connect with others. Using these dictionaries, LIWC compares the words in the provided text with the list of words contained in these internal dictionaries, thus, calculating the match percentage for each of its dimensions.

It is important to note that while a high number of features returned by LIWC can be considered psycholinguistic, as they are used to capture emotional and psychological states and processes (e.g., "posemo" for positive emotions or "anxiety" for anxiety-related words), not all of them fall into this category. Certain LIWC features, such as word count or function words, can be viewed as purely linguistic when used in isolation. However, if these features are paired with others in a set, they can also be considered psycholinguistic. For example, when paired with anxiety words, a higher word count can indicate a certain emotional state. For my experiments, I combine various LIWC features for the purpose of detecting personality relevant information. As a concept rooted in psychology, personality can be indicative of one's psychological state, at least in the way individuals reflect it in communication. Due to this, all LIWC features will be referred to as psycholinguistic for this research, in order to emphasise their role.

Due to the high quantity of LIWC features, I later performed a feature selection approach based on their relationship with the type-based personality models. By doing this, I sought to optimise the approach to my experiments and avoid potential noise in feature set.

Subreddit Participation as Features

Similarly to shaping one's behavioural and communication patterns in face-to-face interactions, personality can also be reflected in one's interactions in online spaces [141]. As such, I found it interesting to examine the possibility that participation in particular subreddits could influence the results of personality prediction. As subreddits are often grouped around a single interest point, this experiment can be seen as an examination of the effect that personality has on interests in particular topics.

When observing all the subreddits individually, connecting interests to particular personalities seems to be complicated. Many Reddit participants frequent each subreddit but have not disclosed any personality-relevant data. To avoid this problem, I focused primarily on measuring the frequency of participation in different subreddits by measuring the number of users for each subreddit and the number of messages posted on them over time.

In order to obtain detailed Reddit information, I used the *PushShift Reddit dataset* [142]. Through it, I have collected participation statistics for different subreddits in the span of time between the chronologically first and last comments present in the PANDORA dataset. After that, 50 most popular subreddits were selected to construct a feature vector using information such as the number of comments posted on the subreddit and the number of participating users over the observed period. Subsequently, these feature vectors were normalised in an effort to apply them to the linear regression models.

5.3.3 Model Selection Approach

After examining the effects of each feature set on personality prediction, I conducted experiments applying different prediction algorithms for this task. To maintain the comparability with the baseline approach, I tested multiple linear regression models. Additionally, I conducted experiments using a deep learning model (KerasRegressor) and an ensemble-learning approach (XGBoost).

Lasso Regression

While the baseline approach relied on the linear regression model implementing the L^2 regularisation norm, due to a large number of n-gram features present in the feature set, a model implementing L^1 regularisation could yield better results [143]. While L^2 regularisation introduces the squared magnitude of coefficients as a penalty function, the L^1 uses the absolute value – making it more robust to outliers. As a result, the L^1 regularisation can impact and potentially eliminate some less important features from the numerous n-grams used in the feature set. A regression model that uses the L^1 regularisation norm is also known as Lasso regression (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator, abbrev.).

The difference between these two models can be mathematically formulated in the following way. If one has m features and n observations in their data, with $x_{i,j}$, they can mark the j-th feature of the i-th observation. Next, if they use w to

represent the weight of their function for the i-th feature of the observation that they are interested in predicting (which I mark with y), the basic regression formula can be written as Equation (5.1). For the L^2 regularisation with the regularisation parameter $\alpha\epsilon[0,1]$ multiplied with the sum of squared weights w, the regression formula takes the form of Equation (5.2). Finally, if one takes the absolute value of weights instead of squaring them, the L^1 regularisation as shown in Equation (5.3) is obtained.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{i,j}w_j)^2 \tag{5.1}$$

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{i,j}w_j)^2 + \alpha\sum_{j=1}^{m} w_j^2 \tag{5.2}$$

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{i,j}w_j)^2 + \alpha\sum_{j=1}^{m} |w_j| \tag{5.3}$$

Elastic-Net

While eliminating less critical features could prove beneficial, a more moderate approach should yield better results. While L^1 regularisation tends to be more strict by removing features, L^2 only minimises their effect. A balanced combination of the two regularisation norms can prove to be beneficial in improving predictions, as it combines the best aspects of both the Ridge and Lasso regression models. The algorithm that relies on both the L^1 and L^2 norm is known as Elastic-Net and can be mathematically formulated in the following way:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} x_{i,j}w_j)^2 + \alpha_1\sum_{j=1}^{m} w_j^2 + \alpha_2\sum_{j=1}^{m} |w_j| \tag{5.4}$$

My theory is that the combination of two regularisations can yield better results as it would simultaneously minimise the effect of the outliers on the prediction while preserving those features that could capture the intricate nature and finer differences between personality traits.

Huber Regressor

Lasso, Ridge and Elastic-Net rely on the ordinary least squares formula for their loss function. One problem with this is that outliers often have too much influence on the predictions. This is true for both the models that implement the L^1 and the L^2 regularisation norms, even though the L^1 norm used the median as the central value to minimise this effect. Several different regression approaches offer the complete elimination of outliers, with one example being RANSAC (**RAN**dom **SA**mple **C**onsensus).

However, due to the size of the data, as well as the nature of the task, I propose that it is best to minimise the effect of outliers rather than eliminate them

entirely. For this reason, I decided to experiment with the Huber regressor, which is available through Python's Sklearn package[5]. Much like the Ridge regression model, the Huber regressor implements L^2 regularisation. However, it does so by using M-estimators [144] rather than the mean of the distribution as its central value, thus making it more resistant to outliers. Due to this property, I speculate that it will result in slightly better predictions than the baseline approach.

The loss function of the Huber regressor can be mathematically formulated in the following way:

$$\min_{w,\sigma} \sum_{i=1}^{n} \left(\sigma + H_\epsilon \left(\frac{x_{i,j} w - y_i}{\sigma} \right) \sigma \right) + \alpha \sum_{j=1}^{m} w_j^2 \tag{5.5}$$

where $H_\epsilon(z)$ takes the values of:

$$H_\epsilon(z) = \begin{cases} z^2, & \text{if } |z| < \epsilon, \\ 2\epsilon|z| - \epsilon^2, & \text{otherwise} \end{cases} \tag{5.6}$$

Epsilon-Support Vector Regression – SVR

Similarly to Huber regressor, the Epsilon-Support Vector Regression, or SVR, has shown good resistance to outliers. Based on a Support Vector Machines (SVM) classification algorithm, SVR uses a kernel trick to perform regression in higher dimensions. As a result, SVR tends to generalise well without its computational complexity depending on the dimensionality of the problem [145]. This generalisation is mainly the result of SVR using an ϵ-insensitive region (also known as an ϵ-tube), that is often used to better approximate functions that have continuous values. With this in mind, as well as the fact that SVR is known to perform well on smaller sets of data, my theory is that the overall prediction scores can be improved by applying the SVR algorithm in my experiments.

Keras Regressor

Deep learning methods have recently shown promising results in the field of automatic personality recognition [8, 126, 66]. While several architectures for deep learning models have achieved promising results, I decided to focus on KerasRegressor – a part of the Keras library, due to its simplicity.

Keras[6] is a high-level library for deep learning in the Python programming language that allows for the easy and efficient construction of neural networks. As part of it, KerasRegressor represents a deep learning model trained to predict continuous values, such as stock prices and weather conditions. In my work, I have experimented with several architectures for the KerasRegressor model, subsequently selecting the best-performing one.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.HuberRegressor.html; last accessed on the 26th of May 2023

[6] https://keras.io/; last accessed on the 26th of May 2023

The model consists of four fully connected layers, with the input shaped to match the data. I use a truncated normal kernel initialiser and ReLU activation function, with Adam as an optimiser. The model is compiled using the root mean squared error as a loss function and trained over 30 epochs with a batch size of 32 to accommodate for the size of data.

Boosting Algorithms

Boosting algorithms are a helpful option when working with weak estimators. Boosting hierarchically builds a model attempting to minimise the error over time. The three most popular ensemble learning models implementing gradient boosting algorithms are XGBoost[7], LightGBM[8] and Catboost[9]. While the first two algorithms utilise asymmetrical trees – with XGBoost growing vertically and LightGBM horizontally, Catboost relies on symmetrical trees. All three algorithms have shown good performances on various prediction tasks.

For this research, I selected XGBoost because it relies on asymmetrical trees that expand level-wise rather than leaf-wise, as well as the splitting method it uses. Additionally, recent works in personality computing have reported promising results when using XGBoost to predict the MBTI personality types [146]. However, it should be noted that boosting algorithms are not advised for smaller data sets or in cases where the features outnumber the data samples, as this can lead to overfitting.

These factors could pose an issue in the case of the PANDORA dataset, as the amount of Big Five labels is relatively small. In order to minimise the risk of overfitting, I used 100 early stopping epochs [94] and performed five-fold cross-validation during the training process.

Since XGBoost uses many different hyperparameters, it is difficult to tell which combination would lead to the most optimal results. For this purpose, I relied on the Optuna[10] package to search the hyperspace for the best possible combination of parameters for my experiments. Table 5.2 lists the parameters and their values calculated by the optimisation package used in my work.

Table 5.2: Hyperparameters used by the XGBoost model for the experiments.

Parameter Name	**Parameter Value**
Number of estimators	10,000
Learning rate	0.002
Maximum depth of a tree	3
L^1 regularisation term	5.25
L^2 regularisation term	34.85
Subsample of columns	0.1
Subsample of training instances	0.7
Gamma	0

[7] https://xgboost.readthedocs.io/en/stable/; last accessed on the 26th of May 2023
[8] https://lightgbm.readthedocs.io/en/v3.3.2/; last accessed on the 26th of May 2023
[9] https://catboost.ai/; last accessed on the 26th of May 2023
[10] https://github.com/optuna/optuna; last accessed on the 26th of May 2023

5.3.4 Ethical Approach to Personality Research

Before going over the results of my experiments, it is essential to address the ethics in personality computing. While the field itself has been rapidly developing [7], one of the most frequently stated reasons for the lack of easily accessible personality-relevant data has been privacy concerns. Due to these reasons, the work of Kosinski et al. [53] on the *MyPersonality* dataset had to be removed from the internet.

A report by Fang et al. [9] stated that only about 10% of research papers reflect on the ethics and fairness of research into personality. This poses an issue of the utmost importance for the research field, as not addressing it can hinder progress of the research field in general. The improper handling of private data can lead to personal information being used in unintended and harmful ways, such as profiling and targeting individuals with particular services or advertisements.

In order to ensure ethical research in my study [147], I complied with the guidelines specified by the Reddit social media platform. Additionally, my study complies with the set of rules specified by the authors of the PANDORA dataset[11] [30]. As a result, I removed data from any user whose information can no longer be publicly accessed through the Reddit platform. Additionally, all findings of my research are reported on an aggregate level only, assuring the protection of privacy for the participants.

5.4 Results

Section 5.3 details several approaches that rely on different features and algorithms. Due to the large quantity of these approaches, and for the sake of providing a detailed and structured comparison between the results of my experiments, I separated this chapter into two subsections. Figure 5.6 provides a general overview of the flow that the experiment process follows. In the first subsection, the focus is on listing the results achieved through the feature selection approach. On the other hand, the second part reports the results for each regression algorithm applied. The results presented in these subsections are summarised and further discussed in Section 5.5.

Before detailing the results of my experiments, I briefly go over the results of the baseline approach and the evaluation criteria. Gjurković et al. [30] were first to test the hypothesis of using MBTI and Enneagram predictions for the prediction of Big Five labels. In their experiments, they used several different feature sets, such as n-grams and MBTI and Enneagram predictions, which were produced by logistic regression models. These features were later combined with different regression methods to acquire the predicted Big Five labels. In Table 5.3, I present the correlations between the predicted MBTI types and Big Five traits with the ground truth labels present in the PANDORA dataset, which was reported by Gjurković et al. [30].

Their best-performing model was an L^2 regularised linear regression model that used a combination of n-grams and predictions of the MBTI/Enneagram la-

[11] https://psy.takelab.fer.hr/datasets/all/pandora/; last accessed on the 26th of May 2023

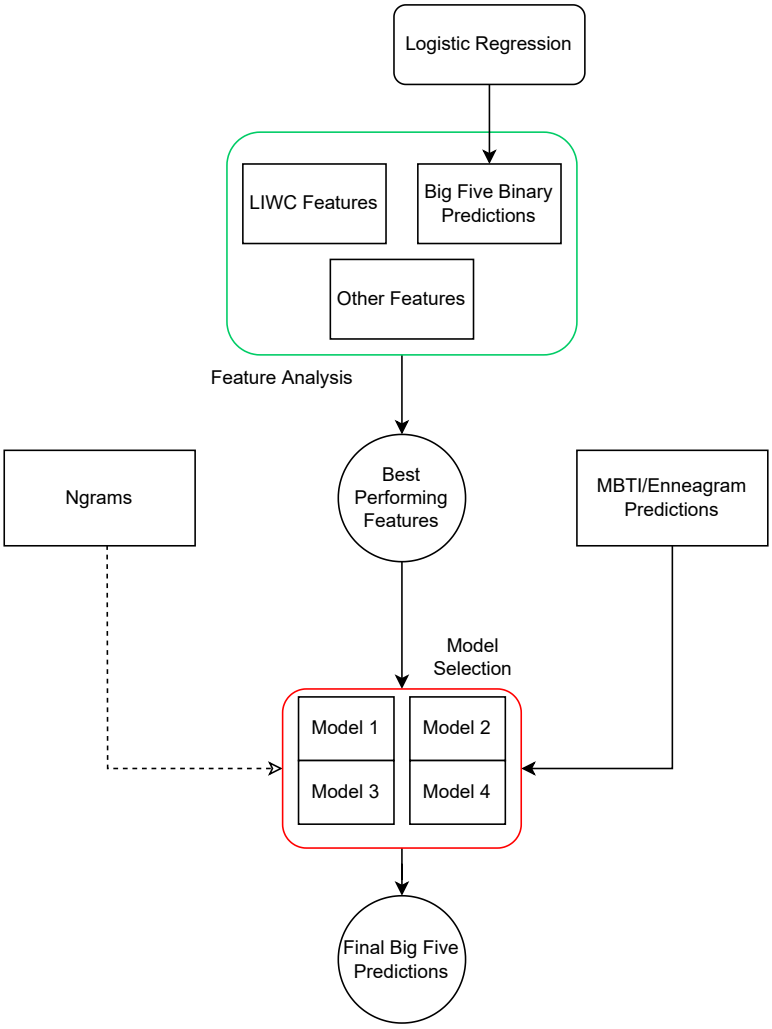

Figure 5.6: Simplified illustration of the pipeline behind my method. The green rectangle depicts feature selection approaches, while the red one highlights the model selection step.

bels as features. This model yielded the best results for nearly all of the Big Five personality traits. The only exception was the *Openness* trait, which demonstrated better performances when the same regression model was used, only without the MBTI/Enneagram predictions in the feature set. The metric used to evaluate these models' performance is the Pearson correlation coefficient [148]. Results of their experiments are reported in Table 5.4 to allow for a comparison with the results produced by the experiments I conducted.

For the sake of readability, when reporting correlation scores for the four MBTI dichotomies, I report a score for only a single value of the two represented by a dichotomy. The reason behind this decision can be explained by the fact that a score for the other value would be equal to the same number multiplied by -1. As each, the value that is not reported in results represents an antipodal point of the reported one. For example, if the *Introverted* value of the *E-I* type shows correlation of 0.125, the correlation of the *Extroverted* value would be -0.125.

Additionally, as the Big Five traits are also known by the acronym **OCEAN** (*Openness*, *Conscientiousness*, *Extroversion*, *Agreeableness*, *Neuroticism*), for visi-

Table 5.3: The results of the baseline approach. The Pearson correlation coefficient scores were adapted from the work of Gjurković et al. [30] CC-BY-NC.

	O	C	E	A	N
Pearson corr.	0.250	**0.273**	**0.387**	**0.270**	**0.283**

Bold indicates best result reported by Gjurković et al. [30].

Table 5.4: The Pearson correlation coefficient between the gold-standard Big Five labels and the predicted values of MBTI types and Big Five traits. Correlations adapted from the work of Gjurković et al. [30] CC-BY-NC.

	O	C	E	A	N
Predicted...					
Introverted	−0.082	0.039	−0.262	−0.003	−0.002
Intuitive	0.127	−0.021	0.049	0.060	0.001
Thinking	−0.001	0.038	−0.039	-0.259	-0.172
Perceiving	0.018	−0.241	0.007	0.034	0.039
Predicted...					
Openness	0.147	−0.082	0.212	0.145	0.070
Conscientiousness	−0.007	0.237	0.013	−0.112	−0.090
Extroversion	0.098	-0.028	0.272	0.044	0.022
Agreeableness	0.006	−0.079	0.023	0.264	0.176
Neuroticism	−0.048	−0.025	−0.042	0.231	0.162

Underlined numbers indicate significant correlation ($p < 0.05$).

bility sake they will be reported by their starting letter in table headers.

5.4.1 Feature Analysis

Big Five Classification Predictions – Median Split

When observing the results of the Big Five predictions achieved through the classification method, one can notice that the overall correlation coefficients, in fact, decreased when compared to the baseline approach (Table 5.6). While the correlations between the predicted Big Five traits treated as classes and the actual Big Five labels seem to be comparable to the results previously reported in the work of Gjurković et al. [30], the predictions of every single personality trait decreased, with the only exception being the *Openness* trait.

The likely reason for this is that the predicted values made by the classification model of the *Openness* trait seem more statistically independent from other personality traits. This is confirmed in the statistical correlation exhibited by predictions which were made by the regression model (Table 5.5).

Table 5.5: Correlations between the gold-standard Big Five labels and predictions that use median-delimited Big Five categories as features.

	O	**C**	**E**	**A**	**N**
Median Preds.					
Openness	0.198	−0.053	0.121	0.005	−0.036
Conscientiousness	−0.078	0.227	−0.039	−0.037	−0.061
Extroversion	0.145	−0.043	0.329	0.012	−0.105
Agreeableness	0.029	0.031	−0.001	0.236	0.171
Neuroticism	−0.008	−0.023	−0.028	0.160	0.246

Underlined numbers indicate significant correlation ($p < 0.05$).

Table 5.6: Pearson correlation coefficient between the actual Big Five traits and the predictions achieved through usage of n-grams, MBTI/Enneagram predictions and median-split Big Five predictions as features.

Features	**O**	**C**	**E**	**A**	**N**
n-grams + Median Split Preds.	**0.260**	0.184	0.336	0.246	0.257
n-grams + Median + Other Preds.	**0.270**	0.225	0.375	0.263	0.255

Bold numbers mark a result that is outperforming the baseline.

Big Five Classification Predictions – Quartile Split

When treating the prediction of the Big Five traits as a three-class classification problem rather than a two-class one, one can note that the correlations between these features and actual Big Five labels were worse than when the median values were used. This can be seen in the results reported in Table 5.7. While *Openness* remains the only personality trait to see improvements over the baseline approach, this is only when MBTI/Enneagram predictions are included in the feature set (Table 5.8).

Table 5.7: Correlations between the gold-standard Big Five labels and predictions that use the quartile-delimited Big Five categories as features.

	O	**C**	**E**	**A**	**N**
Quartile Preds.					
Openness	0.235	−0.025	0.160	-0.011	-0.018
Conscientiousness	−0.071	0.251	0.010	−0.039	−0.078
Extroversion	0.171	0.066	0.350	0.039	−0.011
Agreeableness	−0.011	-0.086	−0.028	0.284	0.169
Neuroticism	−0.039	−0.107	−0.067	0.152	0.234

Underlined numbers indicate significant correlation ($p < 0.05$).

Language-Based Features

Analysing the psycholinguistic features, it should be noted that several researchers in the past have found correlations between the Big Five traits and various

Table 5.8: Pearson correlation coefficient between the actual Big Five traits and predictions achieved through the usage of n-grams, MBTI/Enneagram predictions and quartile-split Big Five class predictions in the feature set.

Features	**O**	**C**	**E**	**A**	**N**
n-grams + Quartile Preds.	0.243	0.222	0.372	0.266	0.246
n-grams + Quartile + Other Preds.	**0.259**	0.258	0.386	0.249	0.265

Bold numbers mark a result that is outperforming the baseline.

LIWC dimensions [48, 113, 114]. However, as the detailed list of LIWC dimensions that correlate with each Big Five trait tends to differ on a dataset-to-dataset basis, it is possible that contextual information, in addition to personality, can have a significant influence on language usage that the software measures.

While it is possible to perform a detailed research into how the Big Five personality traits have influenced language usage on the social media platform Reddit, such a study and its results could introduce information leak into the prediction model if used for the regression task. In addition to this, the usage of psycholinguistic information based on its relationship with the Big Five personality traits could minimise the effectiveness of MBTI/Enneagram predictions present in the feature set.

Instead, I focused on a statistical analysis of the MBTI types and how they influence language use on Reddit, as suggested by data in the PANDORA dataset. Through this approach, I not only open the possibility of this information being leveraged in my prediction model but also provide insight into the linguistic nature of MBTI types. Tables 5.9 and 5.10 include information about all the correlations present between LIWC dimensions and MBTI types.

Table 5.9: Correlation between MBTI types and LIWC features present in the PANDORA dataset. LIWC features that correlate with at least one MBTI type are shown in the table.

	I	**N**	**T**	**P**
LIWC Dim.				
achieve	−0.020	0.050 **	0.091 **	−0.057 **
adverb	−0.010	−0.012	−0.193 **	0.030 **
affect	−0.079 **	0.014	−0.208 **	0.025 *
AllPunc	0.024 *	−0.001	0.013	−0.009
anger	−0.034 **	0.002	0.068 **	0.075 **
anx	−0.027 **	0.009	−0.184 **	−0.038 **
Apostro	−0.019	−0.022 *	−0.100**	0.010
article	0.049 **	0.034 **	0.197**	0.014
assent	−0.057 **	−0.043 **	−0.111 **	0.051 **
auxverb	−0.030 **	−0.002	−0.024 *	−0.004
bio	−0.062 **	−0.031 **	−0.080 **	−0.011
body	−0.053 **	−0.049 **	−0.038 **	0.010
cause	−0.007	0.048 **	0.153 **	0.019
certain	−0.019	0.059 **	−0.046 **	0.014
cogmech	−0.026 *	0.049 **	−0.068 **	0.007
Colon	0.033 **	−0.023 *	−0.038 **	0.009
Comma	0.022 *	0.045 **	0.007	0.002
conj	−0.053 **	−0.012	−0.137 **	−0.029 **
Dash	0.036 **	−0.016	0.024 *	−0.034 **
death	0.061 **	0.037 **	0.088 **	0.069 **
Dic	−0.047 **	−0.012	−0.121 **	−0.047 **
discrep	0.027 *	0.001	0.068 **	−0.026 *
excl	0.003	−0.006	−0.028 **	0.037 **
Exclam	−0.077 **	−0.033 **	−0.211 **	−0.038 **
family	−0.037 **	−0.067 **	−0.105 **	−0.075 **
feel	−0.061 **	−0.015	−0.265**	−0.030 **
filler	−0.040 **	−0.064 **	−0.151 **	0.055 **
friend	−0.118 **	−0.037 **	−0.213 **	−0.020
funct	−0.030 **	0.010	−0.103 **	−0.035 **
future	0.023 *	−0.009	0.108 **	−0.005
health	−0.024 *	0.023 *	−0.069 **	−0.059 **
hear	0.013	−0.022 *	−0.126 **	0.071 **
home	0.007	−0.084 **	−0.051 **	−0.097 **
humans	−0.067 **	0.017	−0.037 **	0.001
i	−0.038 **	−0.076 **	−0.236 **	−0.023 *
incl	−0.087 **	0.001	−0.173 **	−0.076 **
ingest	−0.015	−0.058 **	−0.020	−0.042 **
inhib	0.030 **	0.037 **	0.127 **	−0.047 **
insight	−0.019	0.078 **	−0.067 **	0.020

Note: * when ($p < 0.05$) and ** when ($p < 0.01$).

Table 5.10: Continuation of Table 5.9 – Correlation between MBTI types and LIWC features present in the PANDORA dataset. LIWC features that correlate with at least one MBTI type are shown in the table

	I	**N**	**T**	**P**
LIWC Dim. (2)				
ipron	0.006	0.064 **	−0.057 **	0.026 *
leisure	0.034 **	−0.028 **	−0.033 **	0.051 **
money	0.028 **	−0.014	0.167 **	−0.068 **
motion	−0.036 **	−0.055 **	0.001	−0.095 **
negate	0.035 **	−0.014	0.095 **	0.024 *
negemo	−0.030 **	0.016	−0.017	0.057 **
nonfl	−0.005	0.005	−0.022 *	0.037 **
number	0.054 **	−0.032 **	0.074 **	−0.001
OtherP	0.052 **	−0.004	0.043 **	0.024 *
Parenth	0.047 **	−0.011	0.003	−0.016
past	−0.003	−0.044 **	−0.126 **	−0.025 *
percept	0.001	−0.050 **	−0.224 **	0.022 *
posemo	−0.081 **	0.005	−0.265 **	−0.012
ppron	−0.089 **	−0.064 **	−0.256 **	−0.053 **
preps	0.005	0.045 **	0.002	−0.066 **
present	−0.062 **	−0.012	−0.105 **	0.004
pronoun	−0.070 **	−0.027 **	−0.230 **	−0.033 **
QMark	0.001	−0.019	0.055 **	0.048 **
quant	0.036 **	0.055 **	0.032 **	0.002
Quote	−0.001	0.022 *	0.021 *	0.013
relativ	−0.010	−0.042 **	−0.037 **	−0.075 **
relig	0.038 **	0.041 **	−0.008	0.048 **
sad	0.032 **	0.025 *	−0.121 **	0.001
see	0.043 **	−0.060 **	−0.064 **	0.024 *
SemiC	0.011	0.009	0.021 *	−0.006
sexual	−0.058 **	−0.007	−0.051 **	0.029 **
shehe	−0.072 **	−0.038 **	−0.130 **	−0.036 **
Sixltr	0.014	0.075 **	0.128 **	−0.016
social	−0.127 **	−0.009	−0.145 **	−0.051 **
space	−0.020	−0.008	0.044 **	−0.047 **
swear	−0.049 **	−0.026 *	0.055 **	0.086 **
tentat	0.034 **	0.017	0.001	0.031 **
they	0.011	0.015	0.069 **	−0.045 **
time	0.013	−0.052 **	−0.118 **	−0.059 **
verb	−0.046 **	−0.025 *	−0.127 **	−0.009
WC	0.006	−0.007	0.070 **	−0.023 *
we	−0.066 **	0.030 **	−0.059 **	−0.023 *
work	−0.018	−0.017	0.147 **	−0.068 **
WPS	0.020	0.023 *	0.044 **	0.040 **
you	−0.066 **	0.020	−0.025 *	−0.030 **

Note: * when ($p < 0.05$) and ** when ($p < 0.01$).

The results reported in Tables 5.9 and 5.10 highlight that the two MBTI types that tend to correlate with most of the LIWC dimensions are *T-F* with 69 and *J-P* with 54 statistically significant correlations. To establish whether an entire set of correlating LIWC dimensions can contribute to better predictions of the Big Five traits, I tested all the correlating LIWC dimensions as features, sorting them into four different groups – one for each MBTI type – and a fifth group that includes all 78 LIWC dimensions that have shown correlation with at least one MBTI type. The results of combining these features with *n*-grams and MBTI/Enneagram predictions in the set of features are reported in Table 5.11.

Table 5.11: The Pearson correlation coefficient between the actual Big Five traits and the ones predicted using the combination of *n*-grams, MBTI/Enne. predictions and various LIWC dimensions are divided in sets based on which MBTI type they correlate with.

Features	**O**	**C**	**E**	**A**	**N**
n-grams + ...					
Introverted Correlating LIWC	0.229	0.148	0.321	0.212	0.249
Intuitive Correlating LIWC	0.232	0.159	0.324	0.230	0.238
Thinking Correlating LIWC	0.216	0.150	0.340	0.203	0.241
Perceiving Correlating LIWC	0.228	0.154	0.319	0.216	0.243
LIWC Correlating to All Types	0.214	0.150	0.330	0.206	0.237
n-grams + MBTI/Enne. + ...					
Introverted Correlating LIWC	0.234	**0.274**	0.379	0.258	0.279
Intuitive Correlating LIWC	0.239	**0.283**	0.384	**0.272**	0.266
Thinking Correlating LIWC	0.221	0.273	**0.389**	**0.286**	**0.298**
Perceiving Correlating LIWC	0.235	**0.274**	0.386	0.249	0.282
LIWC Correlating to All Types	0.227	0.271	0.382	0.253	**0.289**

Bold numbers mark a result that is outperforming the baseline approach.

Identifying Useful LIWC Dimensions

While the results of predictions that utilise correlating LIWC dimensions gave promising results for predicting certain traits, it was only when paired with the previously computed MBTI/Enneagram predictions in the feature set that the results improved over baseline. This signals that, despite the LIWC features being an efficient indicator of personality traits, it is only when the relationship between multiple personality models is leveraged that they become most effective.

This is especially apparent for the *Conscientiousness* trait, which had some of the worst results without the MBTI/Enneagram predictions in the feature set, but ended up outperforming the baseline results when predictions of the type-based personality models were reintroduced into the feature set. It is worth noting, however, that due to a large amount of LIWC dimensions correlating with several MBTI types, the potential benefit of certain psycholinguistic features is reduced by the sudden increase in the number of features.

Since this overlap between the LIWC dimensions that correlate with two or more MBTI types ranges from 51.85% shared between the *J-P* and *S-N* types to

86.79% between the *E-I* and *T-F* types, I needed to empirically determine which LIWC dimensions best describe the relationship between MBTI types and the Big Five traits that correlate with them. However, I propose that several factors should be considered to improve the results by helping select the adequate LIWC dimensions for the feature set.

The first factor is too high of a correlation with the MBTI type. While most of the correlations reported in Tables 5.9 and 5.10 tend to be marginally significant, those that have exceeded the absolute value of 0.2 indicate a stronger relationship with the MBTI type and, as such, are not a good indicator of the relationship present between MBTI and Big Five models. The second factor is the degree of correlation between MBTI types themselves. If the LIWC category correlates with both MBTI types that tend to correlate with each other, that LIWC category should be disregarded from the feature set. Finally, the third factor is the relationship between the LIWC dimensions. If the LIWC dimensions correlate with one another, only a single one should be selected, as LIWC categories need to be statistically independent from one another.

Using these three factors as criteria, I was left with the following list of LIWC dimensions that correlate with each of the four MBTI types:

1. **Extroverted**/**Introverted** (**E-I**) type:
   - **shehe** – third person singular pronouns (she, her, him...)
   - **incl** – inclusive words (e.g., with, and...)
   - **number** – numbers (first, thousand...)
   - **present** – present tense verbs (is, does, do...)
   - **posemo** – words associated with positive emotions (love, happy, hope...)
   - **pronoun** – total pronouns (I, they, it...)
2. **Sensing**/**Intuitive** (**S-N**) type:
   - **WPS** – average words per sentence
   - **past** – past tense verbs (walked, were...)
   - **social** – social words (we, thank, care...)
   - **ipron** – impersonal pronouns (that, what, it...)
   - **Colon** – number of colons (:)
3. **Thinking**/**Feeling** (**T-F**) type:
   - **you** – second person singular pronouns (u, yourself, you...)
   - **article** – number of articles (a, an, the...)
   - **sad** – words relating to sadness (:(, cry...)
4. **Judging**/**Perceiving** (**J-P**) type:
   - **Exclam** – number of exclamations (!)
   - **i** – first person singular pronouns (me, myself, I...)

- **hear** – auditory words (hear, sound...)
- **tentat** – tentative phrases (if, any, something...)

Combining these LIWC dimensions with their respective MBTI types, I achieved results in predicting the Big Five traits reported in Table 5.12.

Table 5.12: Pearson correlation coefficient between the gold-standard Big Five labels and predictions achieved through usage of a feature set containing *n*-grams, MBTI/Enneagram predictions and selected LIWC dimensions for each of the MBTI traits (from top to bottom: (1) *E-I*, (2) *S-N*, (3) *T-F* and (4) *J-P*).

Features	**O**	**C**	**E**	**A**	**N**
n-grams + MBTI/Enne. Preds. + ...					
shehe,incl,number,pres.,posemo,pron.	**0.256**	0.270	**0.407**	0.263	**0.296**
WPS, past, social, ipron, Colon	**0.265**	**0.275**	**0.392**	**0.273**	**0.290**
you, article, sad	0.250	0.272	0.381	**0.289**	0.283
Exclam, i, hear, tentat	0.246	**0.283**	0.384	0.269	0.272

Bold numbers mark a result that is outperforming the baseline approach.

The results reported in Table 5.12 indicate that choosing LIWC dimensions with the method I described can further increase the results when predicting Big Five traits, especially in cases when the MBTI type and the Big Five trait have been found to correlate with one another statistically. This is visible for all the MBTI types as the prediction results for *Openness* increased when using LIWC dimensions selected for the *S-N* type, *Agreeableness* increased when using LIWC features for *T-F*, and so on. However, it should be noted that, despite the prediction scores for the *Neuroticism* trait increasing when using LIWC dimensions selected for the *E-I* and *S-N* types, I speculate that this can be attributed either to a possible relationship between the Enneagram and *Neuroticism*. Another explanation is that it is due to the nature of data, as *Neuroticism* is the only trait to follow a bimodal distribution in this dataset (Figure 5.3).

Unlike the MBTI types, Enneagram types have shown no presence of a statistically significant correlation with any of the LIWC dimensions.

Effect of Enneagram Predictions on the Big Five Predictions

While MBTI and Big Five personality models have been previously compared in the works of several authors [10, 103, 129, 72], the relationship between the Enneagram and Big Five traits has not been thoroughly explored. This is primarily because the Enneagram is often underutilised in both academia and consulting – the two areas where the Big Five model and MBTI have enjoyed success. However, taking a closer look into the possible relationship between these models could explain how Enneagram predictions can help predict Big Five personality traits.

The information reported in Table 5.13 indicates that the results change drastically for certain Big Five traits when the Enneagram predictions are removed from the feature set. This is visible when comparing them to the results previously

shown in Table 5.12 as well as the results reported for the baseline approach. With this in mind, I make the following observations:

1. The model's performance when predicting the *Neuroticism* trait without the Enneagram predictions in the feature set decreases in comparison to all the previously used feature sets with Enneagram predictions in them.
2. The model's performance when predicting both *Conscientiousness* and *Agreeableness* increases in almost every case when Enneagram predictions are removed from the feature set. The only exception to this is when the following LIWC dimensions appear in the feature set: shehe, incl, number, present, posemo and pronoun.
3. Predictions of the *Openness* trait either stay the same or only slightly fluctuate when Enneagram predictions are removed from the feature set, indicating that predicting this trait benefits only slightly from Enneagram predictions.

Table 5.13: The Pearson correlation coefficient between the gold-standard Big Five labels and predictions achieved through using *n*-grams, MBTI predictions and selected LIWC dimensions in the feature set with the Enneagram predictions omitted. LIWC dimensions selected for each MBTI type follow the same order described in Table 5.12 (i.e., from top to bottom: (1) *E-I*, (2) *S-N*, (3) *T-F* and (4) *J-P*)).

Features	**O**	**C**	**E**	**A**	**N**
n-grams + ...					
Baseline without Enne. Preds.	0.250	**0.281**	0.374	**0.276**	0.258
n-grams + MBTI Preds. + ...					
shehe,incl,number,pres.,posemo,pron.	0.242	**0.277**	0.380	0.266	0.268
WPS, past, social, ipron, Colon	0.253	**0.285**	0.378	**0.278**	0.267
you, article, sad	0.250	**0.281**	0.370	**0.299**	0.256
Exclam, i, hear, tentat	0.248	**0.293**	0.371	**0.274**	0.248

Bold numbers indicate results that outperform those on the same model that use Enneagram predictions in their feature set.

The first of these three observations points towards a possible relationship between the Enneagram types and the Big Five's *Neuroticism* trait. This can be confirmed when observing the correlations between the Enneagram types and the Big Five traits that Gjurković et al. [30] reported in their work, which I list in Table 5.14 for reference.

Despite these results, it is still difficult to conclude whether the relationship between Enneagram and *Neuroticism* trait results from the data's nature or is because many Enneagram types share the language usage associated with *Neuroticism*. This is primarily due to a lack of literature comparing Enneagram types to Big Five traits, especially from the perspective of language usage. As the PANDORA dataset contains a relatively small number of Enneagram labels, as well as Big Five ones, it would be challenging to conduct an in-depth analysis of the topic from this dataset alone. However, I hope these findings can help to motivate future research into the relationship between Enneagram types, Big Five traits and language usage patterns

shared between them. I conclude that such research would be greatly beneficial for personality computing tasks conducted in the future.

Table 5.14: The Pearson correlation coefficient of the gold-standard Big Five labels with the predicted values of Enneagram types as reported by Gjurković et al. [30] CC-BY-NC.

Features	**O**	**C**	**E**	**A**	**N**
Pred. Type					
Enneagram Type 1	0.002	0.032	−0.028	0.047	0.025
Enneagram Type 2	−0.011	0.108	0.030	0.135	0.046
Enneagram Type 3	0.085	0.014	0.071	−0.064	−0.069
Enneagram Type 4	0.041	0.017	0.033	0.166	0.159
Enneagram Type 5	0.067	−0.035	−0.060	−0.121	-0.076
Enneagram Type 6	−0.051	0.004	−0.035	0.046	0.113
Enneagram Type 7	−0.043	−0.019	0.078	−0.085	−0.088
Enneagram Type 8	0.022	−0.044	0.063	−0.129	−0.075
Enneagram Type 9	−0.034	−0.016	−0.102	0.041	−0.005

Underlined numbers indicate significant correlation ($p < 0.05$).

Subreddit Participation

Analysing data from the PushShift dataset, I found that, in the period between the chronologically first and last comments present in the PANDORA dataset, there has been activity on 879,826 different subreddits. Out of all these subreddits, the 50 most popular ones were those centred around more general topics, such as *r/AskReddit* and *r/worldnews*. However, it is worth noting that several personality-related subreddits were included in the set of most popular subreddits e.g., *r/mbti* and several subreddits dedicated to specific MBTI types, such as *r/INTP* and *r/ENFP*. On the other hand, the PANDORA dataset included information on some 46,214 different subreddits, a considerably smaller number.

After forming feature vectors based on either the number of unique users participating in subreddits within the time window matching that of the PANDORA dataset, or the number of total comments, I found that these two feature vectors are nearly identical. This is because the total number of comments and unique users participating in these subbreddits showed a high Pearson correlation of 0.83 between the two measures. Consequently, I decided to only focus on the feature vector that is formed by using the total number of comments as a way of measuring subreddit popularity. The result predictions which use these features in the feature set are shown in Table 5.15.

While subreddit participation is visibly less effective when predicting *Openness* and *Neuroticism*, it caused a slight increase in the results when predicting *Conscientiousness* and *Agreeableness* with the success rates of predicting *Extroversion* remaining the same. I suggest that this is because many subreddits in the feature vector tend to be more general in nature, rather than topic-specific. This has contributed to them attracting different people who likely do not share much in the way of personality traits. However, as the relationship between interests and

Table 5.15: The Pearson correlation coefficient between the actual Big Five traits and predictions achieved through usage of n-grams, MBTI/Enneagram predictions and subreddit participation in the feature set.

Features	**O**	**C**	**E**	**A**	**N**
n-grams + Subreddits	0.208	0.160	0.331	0.171	0.224
n-grams + Subreddits + MBTI/Enne.	0.225	**0.274**	0.387	**0.274**	0.252

Bold numbers mark a result that is outperforming the baseline approach.

MBTI types is yet to be thoroughly studied, I decided not to further investigate the usefulness of interests in the feature set. Instead, this approach can be left for future works.

5.4.2 Model Selection

Features introduced in the previous subsection were all evaluated on the same L^2 regularised regression model. The same model was previously used as part of a method that achieved the baseline results. While some features led to improvements, they also, in turn, introduced additional complexity in the feature space. To further improve the results, I conducted experiments with several different regression models capable of differently weighing features. These models were tested in hopes of bringing the most out of the features for the task of predicting Big Five traits.

For the sake of conciseness, as well as for easier comparison between the results, I chose to report the results for all the different models within a single table (Table 5.16). For features used as input to these models, I have decided to select the best-performing set, which was a combination of n-grams, MBTI/Enneagram predictions and certain LIWC features. The LIWC features used were selected in a way described in the subsection detailing the method for picking the best language based features (Subsection 5.4.1).

The first section of results in Table 5.16 outlines the results achieved using the Ridge regression model, which is the same as those previously reported in Table 5.12. When comparing these results to other sections of the table, I note that models, such as SVR, Huber regressor and Elastic-Net, led to improvements in predicting most of the Big Five traits. At the same time, Lasso regression, KerasRegressor and XGBoost demonstrated poor performance on the task overall. Out of the better-performing models, Elastic-Net stands out, as it scored the best on three out of five Big Five traits – namely *Openness*, *Conscientiousness* and *Extroversion*. On the other hand, Huber regressor and SVR proved to be better choices for predicting the remaining two Big Five traits (*Neuroticism* and *Agreeableness*, respectively).

The performance of the Lasso regression ended up being the worst-performing model overall. I note that, despite using the L^1 regularisation to remove noise from the feature set, this possibly led to a loss of several important features that were indicative of finer differences between personalities. As personality is a complex concept, it often tends to be affected and manifested through the smallest differences between individuals. As such, I speculate that L^1 regularisation caused the model to be less effective at capturing these slight differences, in turn leading to poor

performances on the task of Big Five personality trait prediction.

KerasRegressor made predictions that correlate slightly worse than the baseline approach across all of the personality traits. While these results indicate worse performances than most of the other models included in Table 5.12, it should be noted that this approach outperformed the BERT-based method on which Gjurković et al. [30] experimented. While I did test different architectures of the KerasRegressor for this task, it is possible that a more complex deep learning model could better capture the relationship between the features used and personality traits.

Table 5.16: Scores for predicting the gold-standard Big Five labels using combinations of *n*-grams, MBTI/Enneagram predictions and LIWC dimensions selected through process described in Subsection 5.4.1 evaluated on different models.

	Features	**O**	**C**	**E**	**A**	**N**
Ridge Reg.	E-I	0.256	0.270	0.407	0.263	0.296
	S-N	0.265	0.275	0.392	0.273	0.290
	T-F	0.250	0.272	0.381	0.289	0.283
	J-P	0.246	0.283	0.384	0.269	0.272
Lasso Reg.	E-I	0.167	0.266	0.358	0.264	0.281
	S-N	0.181	0.268	0.347	0.268	0.270
	T-F	0.170	0.267	0.320	0.256	0.247
	J-P	0.168	0.270	0.327	0.263	0.259
Elastic-Net	E-I	0.269	0.270	**0.408**	0.264	0.310
	S-N	**0.283**	0.283	0.397	0.274	0.298
	T-F	0.263	0.272	0.388	0.289	0.296
	J-P	0.267	**0.285**	0.391	0.265	0.292
Huber Reg.	E-I	0.255	0.269	0.396	0.260	**0.312**
	S-N	0.263	0.276	0.384	0.272	0.288
	T-F	0.245	0.272	0.375	0.284	0.274
	J-P	0.254	**0.285**	0.378	0.268	0.266
SVR	E-I	0.230	0.274	0.370	0.282	0.291
	S-N	0.232	0.267	0.361	0.289	0.286
	T-F	0.242	0.274	0.359	**0.298**	0.279
	J-P	0.242	0.282	0.358	0.294	0.279
Keras Reg.	E-I	0.235	0.179	0.368	0.223	0.228
	S-N	0.234	0.181	0.369	0.220	0.231
	T-F	0.239	0.178	0.359	0.231	0.230
	J-P	0.249	0.171	0.359	0.210	0.227
XGBoost	E-I	0.224	0.219	0.337	0.249	0.285
	S-N	0.216	0.219	0.349	0.253	0.284
	T-F	0.222	0.224	0.335	0.250	0.287
	J-P	0.221	0.217	0.337	0.256	0.286

Underlined numbers outperform the baseline

Bold numbers mark the best performing result.

Similarly to the KerasRegressor model, XGBoost demonstrated less than satisfactory results. Despite improving on predictions of the *Neuroticism* trait over the baseline, the results for other personality traits saw a significant decrease compared to the baseline. Due to the smaller dataset size, I propose that the data-boosting algorithm struggled to correctly predict the right value for each personality trait. Additionally, while the LIWC dimensions differed between feature sets, XGBoost showed almost identical results for each experiment. This leads me to believe that XGBoost is less capable of leveraging this language-related information and instead prioritises other features, such as n-grams and MBTI/Enneagram predictions.

The best-performing solution for the *Agreeableness* trait was the SVR model, which included a subset of LIWC dimensions correlating with the *T-F* MBTI type (e.g., *you*, *article* and *sad*) in its feature set. While not the best-performing model overall, SVR still outperformed the baseline approach using several different feature sets, especially when predicting the *Agreeableness* trait, for which it outperformed the baseline on every single experiment conducted.

I suspect these results are largely due to the SVR's nature to work well with smaller sets of data and due to the error function on which it relies. However, I must remark that SVR had worse results than the baseline on both the *Openness* and *Extroversion* traits – both of which have shown the overall highest correlations with n-grams, indicating that SVR places less importance on these particular features when making predictions.

Another well-performing model is the Huber regressor, which demonstrated overall exemplary performances when predicting the *Conscientiousness* and *Neuroticism* personality traits. Additionally, results for prediction of the other three traits also showed promise. When it comes to the predictions for the *Neuroticism* trait, it can be deduced that the good results are due to the Huber regressor's capability of working well with outliers, as the *Neuroticism* trait has been shown to follow a bimodal data distribution.

Out of all the models, the overall best-performing one seems to be Elastic-Net, which performed best when predicting three of the five Big Five traits – namely *Openness*, *Conscientiousness* and *Extroversion*. The effectiveness of Elastic-Net can be attributed to the good balance of both the L^1 and L^2 regularisation norms, which eliminated noisy, less important, features. This combination of regularisation norms has further helped keep features that influenced personality prediction, thus, utilising them to capture subtle personality differences.

I indicate that, despite the SVR and Huber regressor outperforming Elastic-Net when predicting the *Agreeableness* and *Neuroticism* dimensions, the consistent improvements in scores for many different feature sets point towards Elastic-Net being the best overall choice for the task of predicting Big Five traits with the MBTI/Enneagram predictions present in the feature set.

5.5 Discussion

In the previous section, I outlined and briefly discussed the results from various feature and model-selection approaches. While experiments were conducted

on several different algorithms, including deep learning and ensemble methods, experimentation on linear regression models was more prevalent in highlighting the effectiveness of features while offering better interpretability and comparability with the baseline approach. In the end, the set of best-performing features included MBTI/Enneagram predictions, a set of n-grams stemming from the work of Gjurković et al. [30] and a set of LIWC-based features created through method described in Subsection 5.4.1.

While other features that were used in experiments led to limited improvements in comparison to the best-performing ones, they did offer some insight into the nature of personality traits. Good example of this would be the feature set that treated Big Five personality traits as classes. Although I speculated that it would improve the results overall, as these features would help point the regression model in the right direction, the subpar performance suggests that binning eliminates useful information necessary to make a correct personality prediction (shown in Tables 5.5 and 5.7).

However, this approach still managed to improve the results when predicting the *Openness* trait. These improvements could be attributed to data following a negatively skewed distribution (Figure 5.3). I suggest that this results from the classification predictions for *Openness* being more statistically independent from the other four traits. Still, despite the previous research reporting promising results when using binning strategies [62], I propose that the possibility of information loss vastly outweighs the positives of this approach [149, 150] when predicting personality traits in this manner.

Another set of features I experimented with was Reddit participation and how this reflects on personality. As this particular set of features demonstrated little to no improvement overall, this suggests that personality has little effect on topical interests and how they are expressed through Reddit (Table 5.15). However, as the majority of the most popular subreddits were those that centre around broader topics, it is possible that grouping particular interests into larger classes (e.g., hobbies, music-related and news) and using them as features could lead to a higher correlation with certain personality traits. Due to the breadth of the issue and the overall experimental complexity such a study would warrant, I decided to leave it for future works.

The LIWC-based features introduced in my experiments led to improvements in predicting Big Five traits; however, they also introduced additional complexity in the feature space. An adequate regression model was necessary to best handle this and bring the most out of these features. Algorithms that achieved the best results on each Big Five trait did so when they used LIWC features selected through the methodology described in Subsection 5.4.1. The trait that saw the highest increase was *Openness*, correlating by 0.033 points more, or 13.2% more, with the actual trait values than with the baseline approach.

The results in question were achieved when using a combination of the Elastic-Net model as a predictor and a feature set consisting of n-grams, MBTI/Enneagram predictions and a set of LIWC features selected for *S-N* – an MBTI type with which *Openness* demonstrated a statistically significant correlation in the past. The same model yielded the best results for *Conscientiousness* when the LIWC fea-

tures were selected for the *J-P* type and for predicting *Extroversion* when LIWC features correlated with the *E-I* MBTI type were used.

In the case of these traits, the increase was 0.012 points for *Conscientiousness* and 0.021 for *Extroversion* with a final correlation of 0.408 points for the *Extroversion* trait being the highest correlation value scored on an individual trait. The Huber regressor has proven to be just as successful as the Elastic-Net model when predicting the *Conscientiousness* dimension, as they both achieved the same degree of correlation for this trait.

Additionally, the Huber regressor yielded the best results when predicting *Neuroticism* at 0.312 points correlation, scoring 0.029 points higher than the baseline approach. Finally, SVR model achieved an increase of 0.028 points in correlation over the baseline when predicting the *Agreeableness* trait. For these results, SVR used feature set consisting of n-grams, MBTI/Enneagram predictions and LIWC dimensions selected for the *T-F* type – a type with which *Agreeableness* was found to correlate.

When analysing the results achieved using the deep learning model, it is somewhat surprising to see it not perform as well as the other options. This is especially true when considering the popularity of deep learning approaches for tasks of automatic personality recognition in recent years [8]. However, as Gjurković et al. [30] reported similar results when applying a deep learning algorithm, I can conclude that linear regression models are better choice in leveraging this particular set of features.

This is likely due to the high linearity of these features, as linear regression models are designed to work best in these situations. Another possible contributing factor to ensemble and deep learning approaches performing worse than expected could be that KerasRegressor and XGBoost require more data to be efficient, as the PANDORA dataset is, arguably, smaller in size.

5.5.1 Limitations

While taking a more analytical approach to a study led to many interesting findings regarding different features and models mentioned throughout this chapter, it has also been a double-edged sword. The field of personality computing has been rapidly developing over the last several years, and as such, has seen a number of different approaches applied on the automated personality recognition sub-task. Most of these approaches tend to belong to the deep learning category of artificial intelligence. In order to focus on the linear relationship, I have avoided more complex deep learning methods within this study. However, seeing how they would perform on the same corpora would be interesting.

One additional limitation has been the choice of the evaluation metric. The Pearson correlation coefficient has been primarily selected to allow for the direct comparison with previous research used as a baseline. The significance of this lies in the fact that personality computing has been experiencing issues regarding comparability between different works [7]. Despite this, the information presented in Figure 5.4 which indicates a skewed distribution across all personality traits, suggests that a different evaluation metric would be more suitable for the task (See Chapter 6 for

more details).

5.6 Conclusion

In this chapter, I analysed the effectiveness of different features and algorithms when paired with the MBTI and Enneagram prediction labels on the task of automatic personality recognition. I conducted multiple experiments, testing the performance of several different feature sets and prediction models to explore the relationship between type-based models and the Big Five further.

For my experiments, I looked into the effectiveness of standardising the domain of different personality models by introducing the classification results of the Big Five predictions into the feature set. In addition to this, I also looked into the effect that language features extracted using the LIWC tool have on personality and the effect of social media participation.

The best-performing set of features included MBTI/Enneagram prediction labels, a list of n-grams from previous work and a set of LIWC features selected based on their relationship with the MBTI types. This feature set was then used as input for multiple different regression algorithms as well as a deep learning and a boosting approach. The experiments suggest that an algorithm that utilised L^1 and L^2 normalisation led to the best performance, causing an improvement of up 0.033 points, or 13.2% in correlation strength, for the Pearson correlation coefficient metric on a per-trait level.

One additional experiment I conducted was the analysis of the Enneagram prediction's effect on predicting the Big Five traits (Table 5.13). My analysis indicated that, despite a considerable increase in prediction scores for the *Conscientiousness* and *Agreeableness* traits on several different feature sets, the *Neuroticism* scores were worse every time Enneagram predictions were removed from features. These results signal a possibility of a relationship existing between the Enneagram types and the Big Five's *Neuroticism* trait. This is significant since *Neuroticism* had previously not been found to correlate with any of the MBTI types [10, 102, 103].

Possible directions for this research in the future involve taking a closer look into the effects of interests and topics on personality prediction. While I examined the possibility of subreddit popularity affecting the prediction of traits, such as *Extroversion*, it is possible that specific hobbies and involvement in subreddits centred around them could be a better indicator of one's personality. Additional directions in which this research can be expanded include applying the methods to different data sets. While datasets that include information for multiple personality models are still scarce, other social media platforms, such as Twitter, could prove useful in collecting data for future experiments [78].

Finally, I conclude that the results of this study can be helpful in further understanding personality as they indicate how well it can be captured when translating one set of personality measures to another. The findings of my study can also be beneficial when seeking to create more believable dialogue agents, as it allows for inputs in the form of MBTI personality.

Chapter 6

A Different Way of Evaluating Personality Recognition

In the discussion section of the previous chapter, I have indicated some limitations to the study described within it. One of these limitations has been the choice of the evaluation metric. The present chapter serves as a brief expansion of the previously described research, with the primary purpose being to look into an alternative approach to evaluation, which overcomes the limitations mentioned.

6.1 Introduciton

While Gjurković et al. [30] chose the Pearson correlation coefficient as their evaluation metric, it is essential to note that various metrics have been used in personality computing to evaluate the performances of different models [7]. When focusing strictly on regression problems, the Pearson r correlation is not an uncommon metric. However, the Spearman rank correlation coefficient might be a slightly better choice when the data does not follow a normal distribution. Since the data available through the PANDORA dataset follows largely skewed distribution (Figure 5.3), using Spearman's rank for evaluation could offer more insight into the results.

In addition, Fang et al. [9] have proposed that using more than a single evaluation metric could be a better approach for personality recognition approaches. More specifically, they have mentioned that utilising the information offered by the mean squared error (MSE) in addition to one of the correlation metrics (Spearman's rank or Pearson r) would help better identify the change in error value when making predictions as well as reveal the trend that the prediction model follows.

Throughout the last chapter, I focused on the impact that different features and algorithms have on the task of predicting personality. This was evaluated with the same evaluation metric reported in the approach used as a baseline – the Pearson r correlation. In order use new metrics, I re-did the experiments which the authors of the baseline approach described [30] and evaluated them on Sperman's rank and RMSE (root MSE) metrics. The results of the replicated experiments are presented in Table 6.1.

Looking at Table 6.1 it seems that the results change only slightly when

Table 6.1: Results for the baseline approach. The Pearson correlation coefficient scores were adapted from Gjurković et al. [30] CC-BY-NC; Spearman's rank correlation and RMSE values were manually calculated after replicating the experiments described by the authors of the PANDORA dataset.

	O	**C**	**E**	**A**	**N**
Pearson r	0.250	**0.273**	**0.387**	**0.270**	**0.283**
Spearman's rank	0.250	0.268	0.380	0.283	0.271
RMSE	26.895	29.194	27.889	29.779	30.952

Bold numbers represent the best result reported by Gjurković et al.

comparing Pearson r with the Spearman's rank metric. In fact, on the *Openness* trait, they remain the same. While this is interesting, it should be noted that due to the non-parametric nature of the Spearman's rank and the parametric one of the Pearson r measure, it is difficult to compare the two and draw any solid conclusions.

On the other hand, when observing values for the RMSE, the error seems to be the highest for the *Neuroticism* trait and lowest for the *Openness*. This starkly contrasts the Pearson r measurements, as the scores for *Openness* were reported to be lowest and second highest on the *Neuroticism* for this metric.

6.2 Re-Evaluation of the Previously Described Experiments

Finally, in order to draw a direct parallel with my research described in previous chapter, I have decided to re-run all the algorithms described in Subsection 5.3.3 while using set of features selected through the methodology described in the Subsection 5.4.1. These results are thus directly comparable with the re-evaluated baseline reported in Table 6.1 and with the scores previously mentioned in Table 5.16.

Evaluation on the Spearman's rank correlation coefficient is reported in Table 6.2, while scores for the RMSE are give in Table 6.3.

6.3 Discussion and Concluding Remarks

When comparing the results in Table 6.2 with those in Table 5.16, several interesting observations can be made. Firstly, it seems that improvements over the baseline are more common on average for all models when using Spearman's rank correlation than when using Pearson r. This is especially noticeable in the evaluation of the Keras Regressor, as the improvements are noted on the *Openness* trait for all the features selected. Similar to this, *Agreeableness* has seen improvements for almost all the features selected when using Ridge and Lasso regression, Elastic-Net, Huber Regressor and SVR. I believe this to be the consequence of both the features and algorithms being optimised for the Pearson correlation coefficient rather than for the metrics used for re-evaluation.

Table 6.2: Scores for predicting the gold-standard Big Five labels using combinations of n-grams, MBTI/Enneagram predictions and LIWC dimensions selected through process described in Subsection 5.4.1. Evaluation is conducted using the Spearman's rank correlation coefficient metric.

	Features	O	C	E	A	N
Ridge Reg.	E-I	0.257	0.269	0.400	0.278	0.287
	S-N	0.262	0.272	0.386	0.284	0.285
	T-F	0.249	0.272	0.376	0.295	0.272
	J-P	0.248	0.284	0.378	0.287	0.262
Lasso Reg.	E-I	0.161	0.270	0.368	0.285	0.294
	S-N	0.179	0.273	0.347	0.289	0.278
	T-F	0.164	0.272	0.320	0.260	0.260
	J-P	0.164	0.274	0.325	0.290	0.265
Elastic-Net	E-I	0.269	0.267	**0.403**	0.281	0.303
	S-N	**0.281**	0.276	0.391	0.288	0.292
	T-F	0.264	0.271	0.381	0.296	0.286
	J-P	0.267	**0.286**	0.382	0.285	0.283
Huber Reg.	E-I	0.255	0.267	0.390	0.278	**0.307**
	S-N	0.260	0.270	0.379	0.284	0.285
	T-F	0.246	0.273	0.372	0.291	0.264
	J-P	0.253	**0.286**	0.370	0.287	0.255
SVR	E-I	0.226	0.275	0.371	0.281	0.290
	S-N	0.229	0.271	0.361	0.289	0.280
	T-F	0.238	0.277	0.356	**0.301**	0.272
	J-P	0.239	0.284	0.355	0.297	0.273
Keras Reg.	E-I	0.256	0.171	0.360	0.220	0.226
	S-N	0.254	0.172	0.367	0.218	0.228
	T-F	0.257	0.167	0.356	0.233	0.224
	J-P	0.258	0.162	0.353	0.209	0.221
XGBoost	E-I	0.226	0.220	0.336	0.254	0.280
	S-N	0.219	0.222	0.346	0.257	0.278
	T-F	0.223	0.231	0.333	0.257	0.284
	J-P	0.223	0.218	0.333	0.260	0.281

Underlined numbers outperform the baseline

Bold numbers mark the best performing result.

However, it is important to note that the same models seem to be best performing for the same sets of features selected as in the case of Pearson r correlation evaluation. Elastic-Net was confirmed to give the best results for the *Openness* trait when using features selected for the *S-N* MBTI type. Additionally, the best results were confirmed when predicting *Conscientiousness* and *Extroversion* using features selected for *J-P* and *E-I* types, respectively.

The Huber regressor has once again been found to predict *Conscientiousness* successfully, achieving the same Spearman's rank score as Elastic-Net, using the

Table 6.3: Scores for predicting the gold-standard Big Five labels using combinations of n-grams, MBTI/Enneagram predictions and LIWC dimensions selected through process described in Subsection 5.4.1 for different models, evaluated using the RMSE metric.

	Features	O	C	E	A	N
Ridge Reg.	E-I	26.831	29.223	27.620	29.835	30.831
	S-N	26.781	29.161	27.826	29.754	30.863
	T-F	26.908	29.198	27.921	**29.572**	30.945
	J-P	26.928	**29.099**	27.915	29.789	31.065
Lasso Reg.	E-I	28.650	29.371	28.314	29.803	31.098
	S-N	28.507	29.287	28.499	29.770	31.133
	T-F	28.539	29.259	28.777	30.045	31.678
	J-P	28.651	29.289	28.688	29.886	32.118
Elastic-Net	E-I	26.681	29.454	**27.578**	29.834	30.678
	S-N	**26.564**	29.150	27.747	29.741	30.791
	T-F	26.728	29.438	27.836	**29.572**	30.810
	J-P	26.702	29.124	27.832	29.834	30.835
Huber Reg.	E-I	26.943	29.673	27.767	29.973	**30.659**
	S-N	26.880	29.718	27.923	29.848	30.908
	T-F	27.044	29.638	28.005	29.703	31.082
	J-P	26.960	29.387	27.993	29.903	31.146
SVR	E-I	27.485	30.025	28.510	30.008	30.928
	S-N	27.490	30.091	28.542	29.928	31.045
	T-F	27.416	30.041	28.534	29.771	31.072
	J-P	27.402	29.875	28.591	29.844	31.060
Keras Reg.	E-I	27.118	30.316	28.208	30.498	31.689
	S-N	27.087	30.269	28.162	30.472	31.605
	T-F	27.127	30.418	28.389	30.364	31.569
	J-P	27.033	30.391	28.393	30.530	31.700
XGBoost	E-I	27.024	29.631	28.454	29.968	31.039
	S-N	27.085	29.664	28.425	29.928	31.073
	T-F	27.052	29.630	28.561	29.947	31.018
	J-P	27.064	29.692	28.441	29.882	31.040

Underlined numbers outperform the baseline

Bold numbers mark the best performing result.

same set of features. In addition to this, Huber regressor had the best results for the prediction of *Neuroticism* trait, achieving a score of 0.307. On the other hand, SVR has been confirmed to be the best option for predicting *Agreeableness*, especially when using LIWC features correlating to the *T-F* MBTI type.

Table 6.3 demonstrates that while improvements over the baseline seem less common for the RMSE metric, they offer a better insight into the best choice of algorithm for prediction. While RMSE has decreased only on a couple of test runs for the Huber regressor and SVR, Elastic-Net has seen improvements for almost

every set of features across all traits. In fact, when considering RMSE as a metric, RMSE has been best-performing when predicting *Openness*, *Extroversion* and *Agreeableness*, with the Ridge regression having a slightly lesser error when predicting *Conscientiousness*, and Huber Regressor having smallest error for *Neuroticism* trait.

While using Spearman's rank correlation coefficient has helped confirm the effectiveness of the methods previously tested and evaluated using the Pearson r metric, using RMSE for evaluation has helped single out the best-performing model, which is the Elastic-Net. I believe its success to be largely due to the good combination of the L^1 and L^2 regularisations employed, as was previously mentioned in Subsection 5.5.

Chapter 7

Conclusion of the Thesis

In this chapter, I will briefly reflect on the findings of my research and provide summary of the results and contributions in Section 7.1. After that, in Section 7.2, I will briefly discuss possible directions in which this work could be expanded, effectively providing a conclusion to the thesis.

7.1 Thesis Summary

This thesis covers the broad topic of personality, individual differences and the ways in which they are expressed in textual data. The experiments and analysis conducted throughout the chapters have led to interesting findings that could help further our understanding of the complex concept of personality while also signalling the need for further improvements in some areas. While the first two chapters represent a theoretical introduction to the research areas and topics described within the rest of the work, chapters 3 through 6 describe experimental approaches and introduce subsequent novel findings.

Chapter 3 brings attention to the often neglected utility of textual features for the task of identifying speakers from dialogue utterances. While this research problem has been most frequently tried on audio and signal processing datasets, the transformers were able to predict interlocutors from textual data alone with a certain degree of success. This work is closely related to the work conducted in personality computing, as it establishes grounds for the possibility of individual differences impacting textual transcripts of communication.

Additionally, the research described in the same chapter has resulted in two contributions to the research field of speaker identification. The first of these contributions is a dialogue dataset originating from a commercial video game with over $70,000$ utterances tagged with additional information such as setting and speaker's name. Video games, especially those that are more story-driven, and thus, feature more dialogue, can potentially be useful when creating textual corpora. While storytelling mediums like books and movies are frequently used, I believe that video games are often under-utilised for tasks involving natural language processing.

The second contribution resulting from the same work is a transformer-based method proposed for the speaker identification task. While previous research works

in the field that focused on textual information utilised machine learning and neural network approaches, the approach described in this chapter represents the first time a deep learning model was applied. Subsequently, its application has resulted in a score of over 70% on the F1 metric, significantly outperforming methods used as the baseline.

The next chapter builds on the theoretical implications of the research described in Chapter 3 and describes efforts to understand further what dictates people's communication styles. As such, it details the research into the statistical correlation between personality and various psycholinguistic features present in textual messages used to communicate on the social media platform Reddit. Within this work, personality was observed through dimensions used by two different personality models – the well-established Big Five model, commonly used in personality research and the MBTI model, which has enjoyed great popularity on various social media platforms. Thus, the analysis was able to further substantiate previous research that suggested a relationship between personality measured by the Big Five model and various psycholinguistic features while also discussing novel findings about the possibility of a similar relationship being present between these features and the dimensions used by the MBTI model.

Findings described in the Chapter 4 were then applied in Chapter 5 in an effort to develop an efficient and sustainable framework for translating personality measures from the more easily obtainable ones for the MBTI and Enneagram models into the more scientifically backed Big Five ones. The detailed experimentation approach described in this chapter has subsequently led to interesting insights into the nature of personality and its reflection in textual communication while highlighting the effectiveness of various machine learning models and their capability to capture the linear relationship between personality and different features.

The best-performing approach was found to be one utilising a set of features that includes MBTI/Enneagram prediction labels, a list of n-grams stemming from previous research work and a set of linguistic features originating from LIWC software that were selected based on their relationship with types measured by the MBTI personality model. These features were then coupled with Elastic-Net, SVR and Huber regressor prediction algorithms, showing the most significant improvements on the Pearson r metric – each for different personality traits. Further analysis using the Spearman's rank coefficient and RMSE indicated that the overall best-performing algorithm is the Elastic-Net. This approach has thus led to an increase of up to 13% in the correlation strength between the predicted and actual value for the Big Five personality traits.

The research has additionally pointed towards the possibility of a relationship between the Big Five's *Neuroticism* trait and the types measured by the Enneagram personality model. This is suggested the exclusion of Enneagram predictions from the feature set reducing the accuracy of predictions for the mentioned Big Five trait. The significance of such finding lies in the fact that while the other Big Five traits have been found to correlate with various MBTI types, same has not been confirmed for *Neuroticism*. Thus, the possibility of tying the Enneagram to this trait poses an exciting base for further research.

7.2 Future Work

Due to the broad nature of the research described, there are several different directions in which it can be taken and improved. While each of the subsequent chapters, starting from Chapter 4 seeks to improve upon the limitations of the work discussed in the chapter prior to it, in scientific research there is always room for further improvements.

One such improvement would be further optimisation of the algorithms and features used in Chapter 5 for the evaluation metrics introduced and proposed in Chapter 6. This would not only offer novel insight but also allow for a wider range of works to be able to compare their results with those disclosed in this thesis. As the field of personality computing has had issues with settling on a common evaluation metric, using multiple different criteria of evaluation can be beneficial for future works.

Another improvement could be a further study into the dialogue data from video games, focusing on the personality exhibited by characters participating in dialogues within them. Video games are usually works of fiction. As such, the quality of the dialogue data largely depends on the writer's ability to capture the essence of the personality they assign to their characters. Overly accentuated personality traits, which are more common in literary works, could help better understand properties of prediction models, as they would theoretically require fewer data to be effective.

Finally, the work can be improved by using different features and models, some even more complex. While the research described in Chapter 5 has experimented with some deep learning and boosting approaches, rapid technological advancements continuously create the possibility of testing new algorithms on the same or similar tasks.

With this, I can only hope that the work introduced and discussed in this thesis serves as a solid foundation for future works in the fields of language processing and personality computing, that seek to further our understanding of personality.

Bibliography

[1] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[2] Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, 2020.

[3] Matthew B Hoy. Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[4] Timothy W Bickmore and Rosalind W Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327, 2005.

[5] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, 2014.

[6] Barbara Engler. *Personality theories*. Cengage Learning, 9 edition, 2013.

[7] Le Vy Phan and John F Rauthmann. Personality computing: New frontiers in personality assessment. *Social and Personality Psychology Compass*, 15(7):e12624, 2021.

[8] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339, 2020.

[9] Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel L Oberski. On text-based personality computing: Challenges and future directions. *arXiv preprint arXiv:2212.06711*, 2022.

[10] Robert R McCrae and Paul T Costa Jr. Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57(1):17–40, 1989.

[11] James S Uleman, S Adil Saribay, and Celia M Gonzalez. Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.*, 59:329–360, 2008.

[12] David C Funder. Accurate personality judgment. *Current Directions in Psychological Science*, 21(3):177–182, 2012.

[13] Rajiv Jhangiani, Hammond Tarry, and Charles Stangor. *Principles of social psychology-1st international edition*. BCcampus, 2014.

[14] James Diggle. *Theophrastus: Characters*. Cambridge University Press, 2022.

[15] J. Rusten, I. C. Cunningham, and A. D. Knox. *Theophrastus: Characters*. Harvard University Press, 1993.

[16] Peter N Singer, Philip J Van der Eijk, and Piero Tassinari. *Galen: Works on Human Nature-Volume 1: Mixtures (De Temperamentis)*. Cambridge University Press, 2019.

[17] Jan Strelau and Bogdan Zawadzki. Temperament from a psychometric perspective: Theory and measurement. *The SAGE handbook of personality theory and assessment*, 2:352–373, 2008.

[18] Immanuel Kant. *Kant: anthropology from a pragmatic point of view*. Cambridge University Press, 2006.

[19] Robert M Stelmack and Anastasios Stalikas. Galen and the humour theory of temperament. *Personality and Individual Differences*, 12(3):255–263, 1991.

[20] David E Leary. *Immanuel Kant and the development of modern psychology*. Praeger Publishing, 1982.

[21] Ryan L Boyd and James W Pennebaker. Language-based personality: a new approach to personality in a digital world. *Current opinion in behavioral sciences*, 18:63–68, 2017.

[22] Andrew Brook. *Kant and Freud*, chapter 2, pages 20–39. Palgrave Macmillan UK, London, 2003.

[23] HJ Eysenck and SBG Eysenck. *Eysenck Personality Questionnaire-Revised (EPQ-R)*. Wiley Online Library, 1984.

[24] Heather E. P. Cattell. *The sixteen personality factor (16PF) questionnaire*, chapter 10, pages 187–215. Springer US, 2001.

[25] Michael C Ashton and Kibeom Lee. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review*, 11(2):150–166, 2007.

[26] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[27] Robert R. McCrae. *The Five-Factor Model of personality traits: Consensus and controversy*. Cambridge Handbooks in Psychology. Cambridge University Press, 2009.

[28] Gordon W Allport and Henry S Odbert. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):22–38, 1936.

[29] John M Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990.

[30] Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online, jun 2021. Association for Computational Linguistics.

[31] Isabel Briggs Myers. The Myers-Briggs Type Indicator: Manual (1962)., 1962.

[32] C. G. Jung. *Psychological types: Volume 6*. Princeton University, 1921.

[33] Sanja Štajner and Seren Yenikent. Why is MBTI personality detection from texts a difficult task? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3580–3589, 2021.

[34] Isabel Briggs Myers. *Introduction to Type: A description of the theory and applications of the Myers-Briggs Type Indicator*. Vision Australia Student Support, 1997.

[35] Don Richard Riso and Russ Hudson. *Personality types: Using the Enneagram for self-discovery*. Houghton Mifflin Harcourt, 1996.

[36] Sudhir H Kale and Samir Shrivastava. The Enneagram system for enhancing workplace spirituality. *Journal of management Development*, 22(4):308–328, 2003.

[37] Rebecca J Kimongo Kemboi, Nyaga Kindiki, and Benard Misigo. Relationship between personality types and career choices of undergraduate students: A case of Moi University, Kenya. *Journal of education and practice*, 7(3):102–112, 2016.

[38] John W Lounsbury, Soo-Hee Park, Eric Sundstrom, Jeanine M Williamson, and Anne E Pemberton. Personality, career satisfaction, and life satisfaction: Test of a directional model. *Journal of career assessment*, 12(4):395–406, 2004.

[39] Scott E Seibert, J Michael Crant, and Maria L Kraimer. Proactive personality and career success. *Journal of applied psychology*, 84(3):416–427, 1999.

[40] Alan S Gerber, Gregory A Huber, David Doherty, Conor M Dowling, Connor Raso, and Shang E Ha. Personality traits and participation in political processes. *The Journal of Politics*, 73(3):692–706, 2011.

[41] Gian Vittorio Caprara, Shalom Schwartz, Cristina Capanna, Michele Vecchione, and Claudio Barbaranelli. Personality and politics: Values, traits, and political choice. *Political psychology*, 27(1):1–28, 2006.

[42] André Blais and Simon Labbé St-Vincent. Personality traits, political attitudes and the propensity to vote. *European Journal of Political Research*, 50(3):395–417, 2011.

[43] Andrew Taylor and Douglas A MacDonald. Religion and the five factor model of personality: An exploratory investigation using a canadian university sample. *Personality and individual differences*, 27(6):1243–1259, 1999.

[44] Jennifer Lodi-Smith and Brent W Roberts. Social investment and personality: A meta-analysis of the relationship of personality traits to investment in work, family, religion, and volunteerism. *Personality and social psychology review*, 11(1):68–86, 2007.

[45] Brent W Roberts, Nathan R Kuncel, Rebecca Shiner, Avshalom Caspi, and Lewis R Goldberg. The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological science*, 2(4):313–345, 2007.

[46] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[47] Francois Mairesse, Marilyn Walker, Matthias Mehl, and Roger Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *J. Artif. Intell. Res. (JAIR)*, 30:457–500, 09 2007.

[48] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296–1312, 1999.

[49] Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. Lexical predictors of personality type. In *Proceedings of the 2005 joint annual meeting of the interface and the classification society of North America*, pages 1–16, 2005.

[50] Jon Oberlander and Alastair J Gill. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse processes*, 42(3):239–270, 2006.

[51] Jon Oberlander and Scott Nowson. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 627–634, 2006.

[52] Alastair Gill, Scott Nowson, and Jon Oberlander. What are they blogging about? Personality, topic and motivation in blogs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3, pages 18–25, 2009.

[53] Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist*, 70(6):543–556, 2015.

[54] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

[55] Matti Wiegmann, Benno Stein, and Martin Potthast. Celebrity profiling. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 2611–2618. Association for Computational Linguistics, 2019.

[56] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[57] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. Recognising personality traits using Facebook status updates. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 14–18. The AAAI Press, Palo Alto, California, 2013.

[58] Ben Verhoeven, Walter Daelemans, and Tom De Smedt. Ensemble methods for personality recognition. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 35–38, 2013.

[59] Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. Workshop on computational personality recognition: Shared task. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 2–5. The AAAI Press, Palo Alto, California, 2013.

[60] Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934, 2015.

[61] Wu Youyou, Michal Kosinski, and David Stillwell. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4):1036–1040, 2015.

[62] Cristina Segalin, Fabio Celli, Luca Polonio, Michal Kosinski, David Stillwell, Nicu Sebe, Marco Cristani, and Bruno Lepri. What your Facebook profile picture reveals about your personality. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 460–468, 2017.

[63] Tommy Tandera, Derwin Suhartono, Rini Wongso, Yen Lina Prasetio, et al. Personality prediction system from Facebook users. *Procedia computer science*, 116:604–611, 2017.

[64] Vivek Kulkarni, Margaret L Kern, David Stillwell, Michal Kosinski, Sandra Matz, Lyle Ungar, Steven Skiena, and H Andrew Schwartz. Latent human traits in the language of social media: An open-vocabulary approach. *PloS one*, 13(11):e0201703, 2018.

[65] Ricelli Ramos, Georges Neto, Barbara Silva, Danielle Monteiro, Ivandré Paraboni, and Rafael Dias. Building a corpus for personality-dependent natural language understanding and generation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1138–1145, 2018.

[66] Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. Deep learning-based personality recognition from

text posts of online social networks. *Applied Intelligence*, 48(11):4232–4246, 2018.

[67] Davide Marengo, Cornelia Sindermann, Jon D Elhai, and Christian Montag. One social media company to rule them all: associations between use of Facebook-owned social media platforms, sociodemographic characteristics, and the Big Five personality traits. *Frontiers in psychology*, page 936, 2020.

[68] Marco Cristani, Alessandro Vinciarelli, Cristina Segalin, and Alessandro Perina. Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 213–222. Association for Computing Machinery, New York, 2013.

[69] Sarah Osterholz, Emily I Mosel, and Boris Egloff. #Insta personality: Personality expression in Instagram accounts, impression formation, and accuracy of personality judgments at zero acquaintance. *Journal of Personality*, 2022.

[70] Matej Gjurković and Jan Šnajder. Reddit: A gold mine for personality prediction. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 87–97, 2018.

[71] Xiaodong Wu, Weizhe Lin, Zhilin Wang, and Elena Rastorgueva. Author2Vec: A framework for generating user embedding. *CoRR*, abs/2003.11627, 2020.

[72] Dušan Radisavljević, Bojan Batalo, Rafal Rzepka, and Kenji Araki. Myers-Briggs Type Indicator and the Big Five model-how our personality affects language use. In *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6. IEEE, 2022.

[73] Zhenkun Zhou, Ke Xu, and Jichang Zhao. Extroverts tweet differently from introverts in Weibo. *EPJ Data Science*, 7:1–22, 2018.

[74] Keira Shuyang Meng and Louis Leung. Factors influencing TikTok engagement behaviors in china: An examination of gratifications sought, narcissism, and the Big Five personality traits. *Telecommunications Policy*, 45(7):102172, 2021.

[75] Barbara Plank and Dirk Hovy. Personality traits on Twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 92–98, 2015.

[76] Ben Verhoeven, Walter Daelemans, and Barbara Plank. Twisty: a multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016)/Calzolari, Nicoletta [edit.]; et al.*, pages 1–6, 2016.

[77] Edward Tighe and Charibeth Cheng. Modeling personality traits of Filipino Twitter users. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 112–122, 2018.

[78] Fabio Celli and Bruno Lepri. Is Big Five better than MBTI? A personality computing challenge using Twitter data. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, pages 93–98. Accademia University Press, 2018.

[79] Vimala Balakrishnan, Shahzaib Khan, Terence Fernandez, and Hamid R Arabnia. Cyberbullying detection on Twitter using Big Five and Dark Triad features. *Personality and individual differences*, 141:252–257, 2019.

[80] Denis Eka Cahyani and Anas Falih Faishal. Classification of Big Five personality behavior tendencies based on study field with Twitter analysis using Support Vector Machine. In *2020 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pages 140–145. IEEE, 2020.

[81] Joan-Isaac Biel and Daniel Gatica-Perez. The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2012.

[82] Elisa Bassignana, Malvina Nissim, and Viviana Patti. Matching theory and data with personal–ITY: What a corpus of Italian YouTube comments reveals about personality. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 11–22. Association for Computational Linguistics, 2020.

[83] Thérese Bergsma, Judith van Stegeren, and Mariët Theune. Creating a sentiment lexicon with game-specific words for analyzing NPC dialogue in The Elder Scrolls V: Skyrim. In *Workshop on Games and Natural Language Processing*, pages 1–9, 2020.

[84] Dušan Radisavljević, Bojan Batalo, Rafal Rzepka, and Kenji Araki. Text-based speaker identification for video game dialogues. In *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3*, pages 44–54. Springer, 2022.

[85] Timothy Hazen, Douglas Jones, Alex Park, Linda Kukolich, and Douglas Reynolds. Integration of speaker recognition into conversational spoken dialogue systems. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 01 2003.

[86] Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to Speak and Act in a Fantasy Text Adventure Game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China, November 2019. Association for Computational Linguistics.

[87] Dávid Sztahó, György Szaszák, and András Beke. Deep Learning Methods in Speaker Recognition: a Review, 2019.

[88] Iulian V. Serban and Joelle Pineau. Text-Based Speaker Identification For Multi-Participant Open-Domain Dialogue Systems. In *NIPS Workshop on Machine Learning for Spoken Language Understanding. Montreal, Canada*, 2015.

[89] Amitava Kundu, Dipankar Das, and Sivaji Bandyopadhyay. Speaker Identification from Film Dialogues. In *4th International Conference on Intelligent Human Computer Interaction: Advancing Technology for Humanity, IHCI 2012*, pages 1–4, 12 2012.

[90] Kaixin Ma, Catherine Xiao, and Jinho Choi. Text-based Speaker Identification on Multiparty Dialogues Using Multi-document Convolutional Neural Networks. In *Proceedings of ACL 2017, Student Research Workshop*, pages 49–55, 07 2017.

[91] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[92] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[93] Ye Zhang and Byron Wallace. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 253–263, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[94] Lutz Prechelt. *Early Stopping - But When?*, volume 1524, pages 55–69. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[95] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472, 2019.

[96] Denis Gordeev and Olga Lykova. BERT of all Trades, Master of Some. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 93–98, Marseille, France, May 2020. European Language Resources Association (ELRA).

[97] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy, July 2019. Association for Computational Linguistics.

[98] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and William Phillip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, Jun 2002.

[99] Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. Cost-Sensitive BERT for Generalisable Sentence Classification on Imbalanced Data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134, Hong Kong, China, November 2019. Association for Computational Linguistics.

[100] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

[101] Elias Iosif and Taniya Mishra. From speaker identification to affective analysis: a multi-step system for analyzing children's stories. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 40–49, 2014.

[102] Douglas A MacDonald, Peter E Anderson, Catherine I Tsagarakis, and Cornelius J Holland. Examination of the relationship between the Myers-Briggs Type Indicator and the NEO Personality Inventory. *Psychological Reports*, 74(1):339–344, 1994.

[103] Adrian Furnham. The Big Five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality. *Personality and individual differences*, 21(2):303–307, 1996.

[104] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

[105] David J Pittenger. Measuring the MBTI... and coming up short. *Journal of Career Planning and Employment*, 54(1):48–52, 1993.

[106] Tammy L Bess and Robert J Harvey. Bimodal score distributions and the Myers-Briggs Type Indicator: fact or artifact? *Journal of Personality Assessment*, 78(1):176–186, 2002.

[107] Bruce A Thyer and Monica Pignotti. *Science and pseudoscience in social work practice*. Springer Publishing Company, 2015.

[108] Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009.

[109] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133, 2004.

[110] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, 2011.

[111] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from Twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE, 2011.

[112] Jacob B Hirsh and Jordan B Peterson. Personality and language use in self-narratives. *Journal of research in personality*, 43(3):524–527, 2009.

[113] Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862–877, 2006.

[114] Thomas Holtgraves. Text messaging, personality, and the social context. *Journal of research in personality*, 45(1):92–99, 2011.

[115] Henry Alexander Murray. *Thematic apperception test.* Harvard University Press, 1943.

[116] Mehul Smriti Raje and Aakarsh Singh. Personality detection by analysis of Twitter profiles. In *International Conference on Soft Computing and Pattern Recognition*, pages 667–675. Springer, 2016.

[117] Dušan Radisavljević, Rafal Rzepka, and Kenji Araki. Personality types and traits—examining and leveraging the relationship between different personality models for mutual prediction. *Applied Sciences*, 13(7):4506, 2023.

[118] Clemens Stachl, Florian Pargent, Sven Hilbert, Gabriella M Harari, Ramona Schoedel, Sumer Vaid, Samuel D Gosling, and Markus Bühner. Personality research and assessment in the era of machine learning. *European Journal of Personality*, 34(5):613–631, 2020.

[119] Frank C Worrell and William E Cross Jr. The reliability and validity of Big Five inventory scores with African American college students. *Journal of Multicultural Counseling and Development*, 32(1):18–32, 2004.

[120] Andrea Fossati, Serena Borroni, Donatella Marchione, and Cesare Maffei. The Big Five inventory (BFI): Reliability and validity of its italian translation in three independent nonclinical samples. *European Journal of Psychological Assessment*, 27(1):50, 2011.

[121] Umit Morsunbul. The validity and reliability study of the turkish version of quick Big Five personality test. *Dusunen Adam The Journal of Psychiatry and Neurological Sciences*, 27(4):316, 2014.

[122] Michael K Mount, Murray R Barrick, and J Perkins Strauss. Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology*, 79(2):272, 1994.

[123] Dimitri Van der Linden, Jan te Nijenhuis, and Arnold B Bakker. The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of research in personality*, 44(3):315–327, 2010.

[124] Wiebke Bleidorn and Christopher James Hopwood. Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2):190–203, 2019.

[125] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855, 2013.

[126] Xiangguo Sun, Bo Liu, Jiuxin Cao, Junzhou Luo, and Xiaojun Shen. Who am I? Personality detection based on deep learning for texts. In *2018 IEEE international conference on communications (ICC)*, pages 1–6. IEEE, 2018.

[127] Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. Personality trait detection using bagged SVM over BERT word embedding ensembles. *arXiv preprint arXiv:2010.01309*, 2020.

[128] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, jun 2019. Association for Computational Linguistics.

[129] Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features. *arXiv preprint arXiv:2204.04629*, 2022.

[130] Charles Li, Monte Hancock, Ben Bowles, Olivia Hancock, Lesley Perg, Payton Brown, Asher Burrell, Gianella Frank, Frankie Stiers, Shana Marshall, et al. Feature extraction from social media posts for psychometric typing of participants. In *International Conference on Augmented Cognition*, pages 267–286. Springer, 2018.

[131] Hussain Ahmad, Muhammad Usama Asghar, Muhammad Zubair Asghar, Aurangzeb Khan, and Amir H Mosavi. A hybrid deep learning technique for personality trait classification from text. *IEEE Access*, 9:146214–146232, 2021.

[132] Yang Li, Amirmohammad Kazemeini, Yash Mehta, and Erik Cambria. Multitask learning for emotion and personality traits detection. *Neurocomputing*, 493:340–350, 2022.

[133] Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE, 2020.

[134] Selver Derya Uysal and Winfried Pohlmeier. Unemployment duration and personality. *Journal of economic psychology*, 32(6):980–992, 2011.

[135] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, jun 2007. Association for Computational Linguistics.

[136] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[137] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[138] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, 32(2):74–79, 2017.

[139] Gabriela Ramírez-de-la Rosa, Héctor Jiménez-Salazar, Esaú Villatoro-Tello, Verónica Reyes-Meza, and Jaime Rojas-Avila. A lexical–availability–based framework from short communications for automatic personality identification. *Cognitive Systems Research*, 2023.

[140] Ryan L Boyd and H Andrew Schwartz. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41, 2021.

[141] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2011.

[142] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[143] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[144] Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011.

[145] Mariette Awad and Rahul Khanna. *Efficient learning machines*. Apress Open, 2015.

[146] Mohammad Hossein Amirhosseini and Hassan Kazemian. Machine learning approach to personality type prediction based on the Myers–Briggs Type Indicator®. *Multimodal Technologies and Interaction*, 4(1):9, 2020.

[147] Nicholas Norman Adams. 'scraping' Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of COVID-19. *International journal of social research methodology*, pages 1–16, 2022.

[148] Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, 2018.

[149] Julie R Irwin and Gary H McClelland. Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, 40(3):366–371, 2003.

[150] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine*, 25(1):127–141, 2006.

List of Publications

- **Dušan Radisavljević**, Rafal Rzepka, Kenji Araki. "Personality Types and Traits – Examining and Leveraging the Relationship between Different Personality Models for Mutual Prediction", Applied Sciences 13, no. 7, 2023.
- **Dušan Radisavljević**, Bojan Batalo, Rafal Rzepka, Kenji Araki. "Myers-Briggs Type Indicator and the Big Five Model–How Our Personality Affects Language Use", 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1-6. IEEE, 2022.
- Mateusz Babieno, Masashi Takeshita, **Dusan Radisavljevic**, Rafal Rzepka, Kenji Araki. 2022. "MIss RoBERTa WiLDe: Metaphor Identification Using Masked Language Model with Wiktionary Lexical Definitions", Applied Sciences 12, no. 4, 2022.
- **Dušan Radisavljević**, Bojan Batalo, Rafal Rzepka, Kenji Araki. "Text-Based Speaker Identification for Video Game Dialogues", Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3, pp. 44-54. Springer International Publishing, 2022.
- Milan Čeliković, Bojan Batalo, **Dušan Radisavljević**, Dinu Dragan, Zoran Anišić. "3D Avatar Platform—A Unique Configurator for 3D Figurine Customization", Proceedings of the 8th International Conference on Mass Customization and Personalization in Central Europe MCP-CE (pp. 19-21), 2018.

Acknowledgements

This thesis symbolises not only the culmination of my academic journey within the last four years but also a culmination of a lifelong process of learning that indirectly played part in its formation. As such, it would not have been possible without all the people I interacted with and spent time with throughout my life. I would like to briefly thank them in this section for their contributions.

Firstly I would like to thank the Japanese Ministry of Education, Culture, Science and Technology (MEXT) for granting me the scholarship and, therefore, the opportunity to pursue education in Japan. This gratitude also extends to Hokkaido University, which allowed me to develop new ideas and pursue my research in an environment that stimulates and encourages innovation.

My wholehearted thanks go to Prof. Kenji Araki, who trusted in me enough to allow me to study under his guidance. The importance of his unstinting and continued support, as well as the answers he provided to my questions and the feedback he has given on my findings, could not be overstated.

In addition, I would also like to express my gratitude to Assistant Prof. Rafał Rzepka for his optimistic attitude, valued feedback, unending patience and willingness to help me whenever I felt stuck with my ideas or was second-guessing my choices and decisions during all of the years I spent in Japan.

Special thanks to the members of the review committee: Prof. Yuji Sakamoto, Prof. Miki Haseyama, Prof. Yoshinori Dobashi and Prof. Toshihiko Ito. Thank you for your time in constructing feedback that would help further refine this thesis.

I also want to thank Prof. Ivan Luković for mentoring me during the years I studied at the university in my home country, as well as for trusting in me enough to encourage me to go and study abroad in Japan. This gratitude extends to all the other professors at the University of Novi Sad, from whom I enjoyed learning. Special thanks go to Dinu Dragan, Vladimir Ivančević, Slavica Kordić, Vladimir Dimitrieski, Miroslav Hajduković, Aleksandar Kovačević, Milan Čeliković, Ilija Kovačević and Mila Stojaković for teaching me many valuable lessons.

Additionally, I would like to express my gratitude to all the teachers that dedicated their time and effort to educating me during my years in primary and secondary education. Many thanks to Milan Miladinović, Lidija Štimac, Marija Ratančić, Radoslav Čvorkov, Ljiljana Štimac, Vladislav M. Todorović, Borislav Bradić, Gospa Stanojević, Nevena Ferlan, Dušan Vuksan, Nemanja Ognjanović and Jasmina Rauš, for putting much effort in making the learning process both educational and entertaining, which in turn inspired me to pursue higher education.

While I put a high value on the knowledge I gained thanks to my educators, it is also worth noting that, in the words of Aristotle, *"Without friends, no one would want to live, even if he had all other goods."*. With that in mind, it would be impossible not to mention within this section all of the wonderful people I was fortunate enough to meet and call friends during my lifetime, as thanks to them, I was able to push through on this journey.

With over twenty years of friendship behind us, I can safely say that my life would not be as fortunate and rich if I never met Bojan Batalo. While it is difficult to capture just how much a lifelong friend means to me in a few sentences, I wish to say that I am endlessly grateful for all the help he has provided due to his selfless nature. With his help, I overcome some of the challenges in my youth that could have forever robbed me of any achievement in life. For this reason, I will forever be grateful, as he has changed my life for the better.

Many thanks to Milica Milutinović for being one of my dearest friends and for always putting enough trust in me to ask for advice. Her willingness to listen to my complaints, help when needed and even just gossip with me always cheers me up. Thank you. I truly appreciate you being part of my life.

I would also like to thank Marko Čančar, Maja Mićunović, Nikola Aleksić, Miloš Tepić, Jovana Rašković and Mislav Biličić for staying in touch over all these years, for making me laugh with your jokes and bringing light to my life. Similarly, I am thankful to Una Vojvodić, Ivan Vrsajkov, Maksim Lalić, Radovan Rašković, Jovan Ivanović, Milan Šović, Đorđe Ivanov, Julija Mirković, Janko Zahorec, Bojana Popov, Marko Vještica, Simona Ravić and Danilo Dimitrijević for being dear friends.

My special thanks go to the friends I have been able to make while staying in Japan. To Sami Sinisalmi, Alexander Hallén, Daniel Lysøe, Adam Ismail, Alex Lefebvre, Ferenc Dicső, Joanna Szwoch, Sergey Pavlov, Lincon Souza, Maha Mahyub, Yeldar Toleubay, Kamilla Enikeeva, Kara Dinissa, Filip Vasić, Gytis Mockus, Sim Seung Woo, Monika Matikainen, Paul Kramer, Kyoichi Shida, Evan Blaine, Victor Tideman and Mateusz Babieno. It is thanks to them that a place on the other side of the planet felt more like home.

My biggest thanks has to go to my girlfriend, Georgia Kirkpatrick, for sticking through thick and thin during all the years I have dedicated to my research. While the process has been long and stressful, the kindness, understanding and patience she was willing to provide at the every step have helped fuel my endeavours. Her support made all the obstacles seem surmountable, and she helped me pick myself up whenever I stumbled. With all my heart – thank you.

Last, but not least, I would like to give a huge thanks to my family. To my mother, Dušanka, my father, Dragoslav and my brother Vladislav. Thank you for always showing willingness to help me with my hobbies, despite the distance that separates us. Thank you for showing care and thinking of me, and most of all, thank you for being understanding and supportive of my decision to leave for Japan.

I would like to conclude this section with a final thank you to my grandmother, Višeslava and grandfather, Dragoljub. It is because they always fought tooth and nail for me to get an education that I managed to come this far. To them, I owe my biggest debt, and it is to their memory that I dedicate this entire work.