The Application of Social Media in Modern-Day Influence Campaigns: Personality Profiling and Information Warfare.

Joshua Watt

May 3, 2023

Thesis submitted for the degree of
Master of Philosophy
in
Applied Mathematics / Statistics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

Contents

List of Tables

List of Figures

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: .... Date: 03/05/2023

Acknowledgements

I would firstly like to acknowledge my parents (Vicky Watt and Stuart Watt) as well as my brother (Hamish Watt) who have all been very supportive of me throughout my education. They have always encouraged me to pursue my desired career pathways and have provided me with the necessary guidance over these years, enabling me to obtain Bachelor's and Master's Degrees.

I would also like to acknowledge both my supervisors (Lewis Mitchell and Jonathan Tuke) who have assisted me over the past two years – I couldn't have asked for a better supervision team. They have edited my work, supported me to go to conferences and provided me with general career guidance. I look forward to maintaining a great relationship with them both and I thank them for their constant support.

Finally, I would like to thank two of my teachers: Angela MacDonald (Year 10 Maths) and Philip Moore (Year 12 Specialist Maths) who both helped me realise I was good at mathematics and encouraged me to pursue it at the tertiary level.

Abstract

Social media has become a repository of peoples' information, where nearly 60% of the worlds population share ideas and exchange opinions. While this enables humans to be more connected than ever, it also creates an environment where peoples' data can be used to manipulate opinions at large scales. Our work extends existing techniques which aim to quantify the extent to which public opinion and online discourse can be influenced by companies/governments. Firstly, we explore how an individuals online digital footprint can be used to understand personal attributes about them, such as their personality type. We then consider how this sort of information can be utilised for influence operations in modern-day conflicts, such as the 2022 Russia/Ukraine war.

Personality profiling has been utilised by companies for targeted advertising, political campaigns and vaccine campaigns. However the accuracy and versatility of such models still remains relatively unknown. Consequently, we aim to explore the extent to which peoples' online digital footprints can be used to profile their Myers-Briggs personality type. We analyse and compare the results of four models: logistic regression, naive Bayes, support vector machines and random forests. We discover that a support vector machine model achieves the best accuracy of 20.95% for predicting someones complete personality type. However, logistic regression models perform marginally worse and are significantly faster to train and predict, highlighting that relatively simple models can outperform complex machine learning models. We acknowledge the presence of substantial class imbalance in our dataset and compare a number of methods for fixing the problems encountered with this. Moreover, we develop a statistical framework for assessing the importance of different sets of features in our models. We discover some features to be more informative than others in the Intuitive/Sensory ($p = 0.032$) and Thinking/Feeling ($p = 0.019$) models. While we apply these methods and models to Myers-Briggs personality profiling, they could be more generally used for any labelling of individuals on social media.

The 2022 Russian invasion of Ukraine emphasises the role social media plays in modern-day conflicts, with both sides fighting in the physical and information environments. There is a large body of work on identifying malicious cyber-activity, but less focusing on the effect this activity has on the overall conversation, especially with regards to the Russia/Ukraine Conflict. Here, we employ a variety of techniques including sentiment/linguistic analysis and time series analysis to understand how certain bot activity influences wider online discourse. In our results we observe that self declared bots most strongly increase discussions of work/governance ($p = 3.803 \times 10^{-18}$) with the most prominent effects after five hours. Moreover, we observe that self declared bots increase angst in the online discourse ($p = 2.450 \times 10^{-4}$) and discussions of motion ($p = 7.93 \times 10^{-10}$) with the most prominent effects after seven hours and three hours, respectively. Discussions of motion were most often involved with staying/fleeing a country and hence self-declared bots were likely influencing peoples' decision to flee their country or not. Our work extends and combines existing techniques to quantify how bots are influencing people in the online conversation around the Russia/Ukraine invasion. It provides a statistical framework which can be applied more generally to any influence campaign on social media and enables researchers to quantitatively understand what makes these campaigns impactful.

Chapter 1

Introduction

Social media is an interactive technology that facilitates the creation and sharing of information through virtual communities and networks. Social media is believed to have existed since the 1960s where the Programmed Logic for Automatic Teaching Operations (PLATO) System was launched, the first online chat room [77]. However, it wasn't until after the launch of Facebook in 2004 that media became available to everyone with internet access, and its usage consequently soared. When combined with the introduction of smart phones and high speed internet, social media has enabled humans to be more connected than ever. Contacting overseas friends or relatives is now possible from anywhere in the world, all with the touch of a button.

Nowadays there are thousands of social media applications and over 4.59 billion people use social media worldwide, constituting approximately 60% of the world's population [46]. While this enables most of the world to be connected, it also creates an environment of mass data, defining what we refer to as the information environment. There are two important aspects of social media which are unique to this type of information environment. Firstly, there are huge amounts of individual-level data that each user provides, and secondly, there is an underlying dialogic transmission system; with many sources of information and many receivers. Consequently, it is crucial for scholars to understand how these two aspects of social media impact society. In what follows, we discuss how these two aspects of social media create a climate which can be weaponised by governments and other organisations.

Every time a user enters a social media application, they leave a unique trace of data – this includes information they have posted, liked, shared, commented and even how long they have spent viewing different material on the application. We refer to this unique trace of data as a user's online digital footprint. It has been suggested that someone's online digital footprint can be used to expose a lot of information about them; including their personality profile, relationship status, political opinions and even their propensity to adopt a particular opinion [162, 111, 143, 144]. One example of this is the British political consulting company, Cambridge Analytica, which were suggested to use peoples' online digital footprints to impact the result of the 2016 US election and the 2016 Brexit referendum by former employee and whistleblower, Christopher Wylie. However, the extent to which companies like Cambridge Analytica can determine this information from social media data is still questioned by many scholars [111, 143, 144]. As a result, it is of interest for many individuals to understand the extent of information that is attainable from their online digital footprint. If companies with access to this data can accurately predict personal information about people, then it is possible for it to be misused and potentially weaponised. This is of key concern for governments, who seek to maintain democracies and the ethical use of such data, both of which can be abused by understanding such personal information.

The dialogic transmission system which underlies social media means there are many sources and receivers of information. This differs greatly from traditional media, such as newspapers, TV and radio, where there are few sources of information but many receivers. Because of this, it is very hard to distinguish between information that is accurate and information that is not, such as misinformation/disinformation. This makes social media a critical tool in information warfare, playing a considerable role in the 2022 Russian/Ukraine war [29, 119], for example disinformation and more generally *reflexive control* [148] have been used by Russia and other countries against their enemies and internally for many years [49]. A relative newcomer in this space – Twitter – has already been extensively used for such purposes during military conflicts, for instance in Donbass [49], but its role in conflicts is evolving and not fully understood. Both sides in the Ukrainian conflict use the online information environment to influence geopolitical dynamics and sway public opinion. Russian social media advances narratives around their motivation, and Ukrainian social media aims to foster and maintain external support from Western countries, and promote their military efforts while attempting to undermine perceptions of the Russian military. Examples of these narratives include allegations: that Ukraine was developing biological weapons [161], that President Volodymyr Zelenskyy had surrendered [28, 85], and that there is a sustained campaign showing the apparent success of 'The Ghost of Kiev' – a mythical MiG-29 Fulcrum flying ace credited with shooting down six Russian planes over Kyiv [89]. Some of the information being pushed is genuine, and some is malicious. It is not easy to discriminate which is which.

As a result, it is important to understand and measure the extent to which the huge amounts of individual-level data and the dialogic transmission system can be utilised by individuals and groups of individuals. We explore this by developing statistical frameworks which underpin the vulnerabilities in these unique aspects of social media. Our analysis is separated into two parts which address: (i) what personal information can be learnt about individual accounts, and (ii) whether groups of automated online accounts can influence many human accounts. We perform an analysis of Twitter data in each case. Firstly, we seek to determine how informative someones' online digital footprint is in predicting their personality type. The Myers-Briggs personality model is a theoretical model comprised of four traits/dichotomies – this model was developed by American personality researchers and based on the theory of Carl Jung [18, 79]. Modelling personal information about individuals using their online information has previously enabled researchers to understand the accuracy of such models. In our research, we extend this work by creating a new labelled dataset of Myers-Briggs personality types on Twitter and a statistical modelling framework which can be generally applied to any labelled characteristic of online accounts. In essence, our work aims to reconsider the personality profiling and political microtargetting performed by companies like Cambridge Analytica – we do this by validating the performance of these types of models and we quantify the importance of their features. Secondly, we seek to discover how influential automated online accounts (bots) are in the online discussion of the Russia/Ukraine war. Measuring and interpreting various language features has previously allowed researchers to understand community dynamics and identify inauthentic accounts and content [140, 120, 9]. Here we apply and extend these techniques to understand and quantify the influence of bot-like accounts on online discussions, using Twitter data focussed on the Russian invasion of Ukraine. In essence we seek to determine whether the malicious influence campaigns work as intended.

In Chapter 2, we provide a background of our work which provides the context and purpose of the research. Firstly, we motivate our analysis by discussing how social media creates an environment that can be exploited by various governments and companies. We then provide an overview of two different personality models: the OCEAN personality model and the Myers-Briggs personality model. The former was created using a statistical approach and the latter was formulated using a theory-based approach. In our analysis, we use the Myers-Briggs personality model. We then provide an overview of the Natural Language Processing (NLP) tools used throughout our research. This includes Linguistic Inquiry and Word Count (LIWC; pronounced "Luke") [113], Valence Aware Dictionary for Sentiment Reasoning (VADER) [74], Bidirectional Encoder Representations from Transformers (BERT) [41], and Botometer [163], a supervised machine learning classifiers which distinguishes bot-like and human-like accounts. We then provide a mathematical background which consists of: the binary models used in to profile the personalities of Twitter users, the statistical methods for hypothesis testing, and data manipulation methods used throughout our analysis. Finally, we perform a literature review of personality profiling work to date as well as work on the detection/influence of bots. As part of this, we provide a detailed overview of the performance of the personality profiling models which utilise the OCEAN and Myers-Briggs frameworks. We find that most of the work in this field focuses on obtaining models of high accuracy and often doesn't acknowledge class imbalances in the data. Hence, we argue for more research focusing on more interpretable models and dealing with class imbalances in data of this type – both of which we explore in this analysis. On the topic of bot influence/detection, we find that a lot of the work focuses on the detection of automated accounts, with very little work focusing on their influence and no work (to our knowledge) focusing on their influence during the 2022 Russia/Ukraine war. This highlights the importance of our research objectives and emphasises the relevance of our results.

In Chapter 3, we aim to determine how informative someones' online digital footprint is in predicting their Myers-Briggs personality type. We do this by first collecting a labelled dataset of accounts with their Myers-Briggs personality types. We observe that people self-report their personality types on Twitter and exploit this by querying for profiles which have done so – these accounts then form our labelled dataset. We collect a number of different features for these accounts including social metadata features and linguistic features. Linguistic features include LIWC, VADER, BERT and Botometer features. A number of preprocessing steps are then performed on the data to ensure it is appropriate for modelling. We then perform an exploratory data analysis (EDA) on the dataset, where we firstly consider any potential biases that may arise as well as the balance of the dichotomies. We find that some of the dichotomies are very unbalanced, which leads us to consider five weighting/sampling techniques in our models. As part of the EDA, we then consider the independence of the dichotomies as well as the various features in our models. We find that the dichotomies are fairly independent, so we will consequently perform independent models on each of the four dichotomies. Moreover, we find that some of the features have high correlations with one another, leading us to perform a principal component analysis, which both reduces the dimension of the feature space and the multicollinearity of the features. Using these features, we then perform four independent logistic regression models on each dichotomy to model the the Myers-Briggs personality type of the accounts. As part of this, we also consider five different weighting/sampling techniques to adjust for class imbalances. We compare the results of the logistic regression models with naive Bayes classifiers, support vector machines and random forest models, and discover that synthetic minority oversampling technique (SMOTE) performs poorly with naive Bayes. Hence, we perform a low-dimensional example of SMOTE with naive Bayes which outlines a discrepancy with combining both these techniques. Lastly, we provide a statistical framework for analysing the importance of different features in these models. We consider the importance of features at an individual level and across groups of features for each dichotomy. As a whole, this chapter outlines how an environment of mass data, like social media, can be used to profile personal characteristics about individuals at large scale.

In Chapter 4, we seek to discover how influential automated online accounts (bots) are in the online discussion of the 2022 Russia/Ukraine war. We do this by first collecting a dataset of Twitter content related to the Russia/Ukraine war over the first two weeks since Russia invaded Ukraine. We queried hashtags in support of Russia/Putin and Ukraine/Zelenskyy to obtain the relevant content. We then performed preprocessing on the dataset which consisted of calculating the Botometer results for a portion of these accounts sharing these hashtags. Moreover, we use these hashtags to calculate a national ‘lean’ for the accounts, which outlines whether the accounts were in support of Russia/Putin, Ukraine/Zelenskyy or a combination of both. We then consider a time series of the bot activity and observe how this aligns with the hashtag activity as well as a number of significant events which occurred over the first two weeks of the war. Next we analyse the distribution of bot probabilities based on the national ‘lean’ of the accounts to examine how each side of the conflict utilise bots. We then consider the effects of bots on the overall discussion surrounding the conflict. We do this by formulating a statistical framework which allows us to observe the effects of bot activity on the linguistic content of the discussion. We consider linguistic features of the discussion such as conversations of angst, friends and motion. As part of this, we discover the types of words which are most frequently occurring for each linguistic category. Altogether, this chapter allows us to discover how the dialogic transmission system underpinning social media can be weaponised by governments in a modern day conflict to influence public discussion at a large scale.

Our work aims to extend existing techniques to understand how companies and governments can utilise social media to profile personal characteristics about individuals and influence public discussion on a large scale. The main contributions are:

- A labelled dataset of approximately 44,000 Twitter users along with their Myers-Briggs personality types, this dataset is the largest available dataset (to our knowledge) of labelled Myers-Briggs personality types on Twitter.
- A statistical framework which combines NLP tools and mathematical models to model/predict the personality type of users online – this same framework can be more broadly utilised to model any labelled characteristics about online accounts.
- A comparison of different machine learning models on NLP features as well as a comparison of various weighting/sampling techniques to address problems with class imbalance.
- A visual low-dimensional demonstration of why SMOTE performs poorly when combined with naive Bayes.
- Statistical methods which compare the importance of different features in NLP-based models at an individual level and across groups of features.
- A dataset[1] of approximately 5.2 million posts created by Twitter users who participated in discussions around the Russian Invasion of Ukraine [158].
- An analysis of the effect which bot activity has on emotions in online discussions around the Russia/Ukraine conflict.

[1] Dataset available at `https://figshare.com/articles/dataset/Tweet_IDs_Botometer_results/20486910`.

- A statistical framework which can be applied to measure how people get influenced in online networks. This framework can be more generally utilised in political campaigns, dis/misinformation campaigns or any online advertisement campaign.

Chapter 2

Background

2.1 Motivation

Social media is a rich information environment[1] where users share experiences, opinions and ideas in virtual communities and networks. As a result, social media has fundamentally changed the way humans consume their information. Online networks like these provide a climate where any sort of information can exist, and the difference between the truth and falsehood is more blurry than ever [108]. While social media allows the world to be more connected than ever, it also has the potential to be misused and abused by a number of different entities. In particular, several companies and governments utilise the rich information environment to manipulate and influence the opinions of people [142, 33, 138, 124, 98]. These entities are motivated by several factors and any form of advertising is always intended to influence opinions [142]. Social media serves as a platform for them to manipulate human opinion at large scale, giving them enormous power. Just imagine if you could convince people that the earth is flat or that your favourite politician should be elected. Social media is an environment where entities can push narratives like these, as misinformation[2] and disinformation[3] is often more attractive than the truth [162]. We explore the extent to which these companies can influence individuals through two different methods. First, we consider the extent to which users' personalities can be profiled through their online digital footprint. Then we consider the extent to which bots influenced the online discussion during the Russian invasion of Ukraine.

Personality profiling has been utilised for decades, as humans often want to better understand themselves and how they interact with their environment. Since the introduction of Social Media, personality profiling has played an important role in political campaigns,

[1] The aggregate of individuals, organizations, and systems that collect, process, disseminate, or act on information.

[2] Incorrect or misleading information which is not deliberately deceptive.

[3] False information that is spread deliberately to deceive people.

digital marketing and employment [111]. In the past, people have obtained their personality profile through undertaking a questionnaire. However, social media has enabled data scientists to model peoples' personality profiles, giving them the ability to profile users without them even knowing it. Cambridge Analytica (CA) was a British political consulting company who utilised personality profiling for what they described as 'behavioral micro-targeting' [68]. Former CA employee and now whistleblower, Christopher Wylie, claimed the company was responsible for using personality profiling to influence the result of the 2016 US Election, the 2016 Brexit referendum and many other political events [162]. He said that CA determined users who were neurotic and more subject to being influenced, they then targeted these users with campaigns specific to their personality. Other authors have questioned and criticised the extent to how accurate CA's models were [142]. While CA's operations were closed in 2018, there are still a number of active entities who are believed to misuse social media data in the same way that CA did [162]. A number of academics have explored the accuracy of predicting peoples' OCEAN[4] personality types, however there has been limited research utilising the Myers-Briggs personality model [12, 144, 143]. Moreover, a majority of these academics have just explored the accuracy of these models, with very limited research which quantitatively evaluates how personality types utilise language differently. As a result, we explore how accurately peoples personality profiles can be modelled through their online digital footprint. We produce a labelled dataset of 43,977 Twitter users, the largest personality labelled Twitter dataset which we know of.

Much of the research concerning bots on social media has involved their detection, with little research measuring their influence [109, 38]. However, with the war in Ukraine being labelled as the 'first to introduce a new front line – the internet' by the BBC, the presence and influence of malicious online campaigns has been of interest for many defence organisations and governments [155, 83]. The use of bots by Russian authorities has been widely observed: *e.g.,* Collins [33] found 5,000 bots were pushing protests against *Russiagate haux,* a political event concerning relations between politicians from US and Russia; and Shane [138] suggested Russia created 'Fake Americans' to influence the 2016 US election. Moreover, Purtill [124] found that Russia had a massive bot army in spreading disinformation about the Russia/Ukraine conflict, and Muscat and Siebert [98] have suggested that both Ukraine and Russia are utilising bot armies in their cyber warfare. However, the extent to which these bots drive particular discussions and influence the behavior of humans on social media during the Russia/Ukraine conflict is relatively unexplored. We aim to address this question through analysing how bots influence topical discussion during the first two weeks of the Russian invasion of Ukraine. We create our own dataset of 5,203,746 tweets and provide a statistical framework for how the influence of bots can be measured – this is generalisable to any bot campaigns on Twitter.

[4] A personality model which measures peoples Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism, forming the OCEAN acronym.

2.2 Personality Models

In this section, we discuss two popular personality models; the OCEAN Personality Model and the Myers-Briggs Personality Model. While we primarily utilise the Myers-Briggs personality model in our results, we also overview the OCEAN personality model as it is used in a majority of the past literature. The OCEAN personality model also has considerable links with the Myers-Briggs model (see Table 2.4) and serves as a good introduction to the Myers-Briggs model. In this Section, we present the psychological framework of both models and consider their development/history. Moreover, we discuss their application and consider their limitations.

2.2.1 OCEAN Personality Model

The OCEAN personality model (sometimes referred to as the Big 5 personality model) is a taxonomy/grouping for personality traits which was created in 1949 by Donald Fiske. The initial model was not popular among academics and was consequently expanded upon by a number of independent researchers including Norman (1967), Smith (1967), Goldberg (1981), McCrae/Costa (1987) and Digman (1990) [150, 43]. The model then began gaining popularity in the 1980's due to the advances by Goldberg and Digman [61]. This model is fundamentally based upon a dataset created by Gordon Allport and Henry Odbert in 1936, where they connected around 4,500 verbal descriptors to certain personality traits [154]. The data was created by scrupulously examining words from the 1925 edition of Webster's New International Dictionary and categorizing all words that appear to refer to human traits [75]. The relationship between the verbal descriptors and the personality traits in the dataset was significantly reduced through utilising factor analysis. In the 1940's, Raymond Cattell and his colleagues narrowed down Allport and Odbert's dataset down to sixteen personality traits. However, numerous independent psychologists including Norman, Smith, Goldberg and McCrae/Costa all found these psychological traits could be further reduced to five main traits [31]. These five main traits then formed a basis for what we know refer to as the OCEAN (or Big 5) personality model. As a result, the OCEAN personality model is a five factor model of personality.

While Norman, Smith, Goldberg and McCrae/Costa all found there to be five dominant traits, they did not necessarily agree on the definitions or names for these traits. However, all traits associated with each of these five dimensions of personality have been found to be factor-analytically aligned and highly inter-correlated [27, 2]. The most accepted labels for these five personality traits are: Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. The model gets its name by taking the first letter of each of these traits to form the acronym OCEAN. Firstly, Openness (or openness to experience) is a general appreciation for adventure, art, unusual ideas, curiosity and variety of experience. People with high openness are more open new new experiences, are more creative and are more aware of their feelings [5]. Conscientiousness is the preference to act dutifully, display self-discipline and strive for achievement. People with high conscientiousness are often prepared, pay attention to details and like order in their lives [35]. Extroversion is classified as a pronounced engagement with the external world. People who are more extroverted are more likely to be the life of the party, feel more comfortable around people and like being the centre of attention [35]. The agreeableness trait aims to reflect differences in general concerns for social harmony. People who are more agreeable are more interested in people, sympathize with others' feelings and take time out for others [131]. Finally, Neuroticism is the tendency to experience negative emotions such as anxiety, depression or anger. People who are more neurotic are more likely to be emotionally unstable, get irritated more easily and frequently worry about things [35]. Each of these five traits are represented on a numerical scale with scores usually being normalized between zero and one hundred [154]. Figure 2.1 describes the OCEAN personality model and gives an overview of what cognitive processes each of these traits encapsulate.

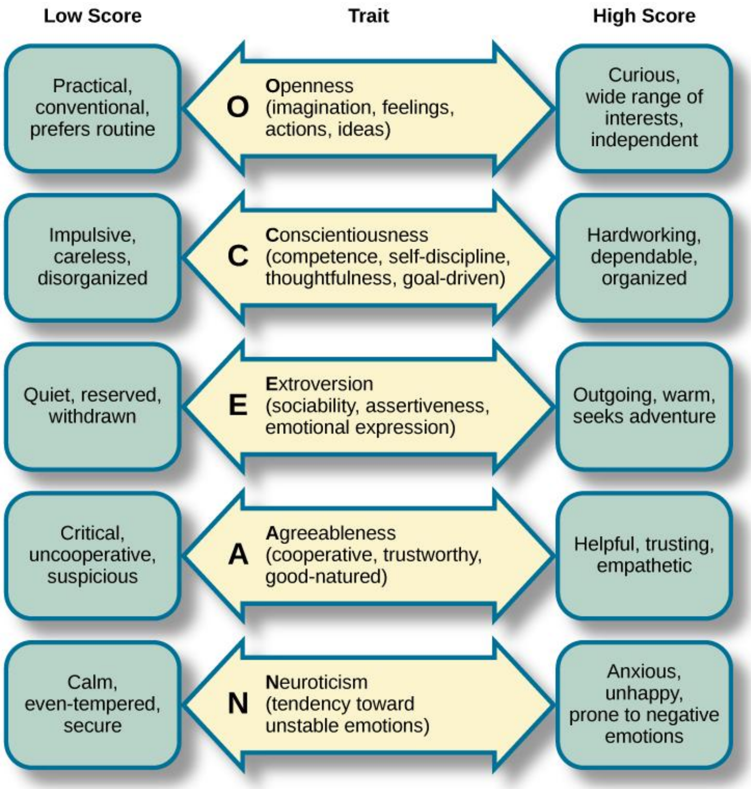

Low Score
Trait
High Score
Practical, conventional, prefers routine
O Openness (imagination, feelings, actions, ideas)
Curious, wide range of interests, independent
Impulsive, careless, disorganized
C Conscientiousness (competence, self-discipline, thoughtfulness, goal-driven)
Hardworking, dependable, organized
Quiet, reserved, withdrawn
E Extroversion (sociability, assertiveness, emotional expression)
Outgoing, warm, seeks adventure
Critical, uncooperative, suspicious
A Agreeableness (cooperative, trustworthy, good-natured)
Helpful, trusting, empathetic
Calm, even-tempered, secure
N Neuroticism (tendency toward unstable emotions)
Anxious, unhappy, prone to negative emotions

Figure 2.1: Personality types in the OCEAN personality model [64].

A method which would target each of these personality traits was subsequently required once researchers in this area had discovered the presence of the five personality traits. Measuring the five OCEAN personality traits is predominately done through the use of self-report questionnaires [47]. The most frequently used measures of these personality traits utilise self-descriptive sentences, however some short forms of the test have been developed when respondent times are limited [149]. While self-report questionnaires are the primary tool used in most personality models, they are always subject to confirmation bias and the Barnom Effect. Note that the Barnom Effect is where people give a high rating to a positive description that supposedly apply specifically to them [116]. Some examples of websites providing the Big 5 Personality Test are: `www.truity.com`, `www.openpsychometrics.org` and `www.bigfive-test.com`. These tests use a combination of questions which are positively correlated with the targeted trait and negatively correlated with the targeted trait. For example, in the case of extroversion, some of the self-descriptive sentences would target the extroverted trait and some self-descriptive sentences would target the introverted trait (we call these reversed statements). The self-descriptive sentences which are present in most OCEAN personality tests can be viewed in [97].

The OCEAN personality model has numerous applications in education, employment, romantic relationships, political identification and religiosity. One major study showed that GPA/exam performance are highly correlated with conscientiousness and academic success is negatively correlated with neuroticism [86]. Another study found that the openness personality trait had a positive relationship with academic achievement in a distant setting – something that would have played a crucial role during the Covid-19 pandemic [152]. In regards to an employment setting, one study found that: openness is positively correlated with higher proactivity at individual and organizational levels, agreeableness is negatively correlated with individual task proactivity, extroversion is negatively correlated to individual task proficiency, conscientiousness is positively correlated to all forms of work performance and neuroticism is negatively correlated with all forms of work performance [78]. Another study aimed to understand how each of the OCEAN personality traits impacted relationship quality in dating, engaged and married couples. The authors found that openness, agreeableness and conscientiousness were all positively correlated with relationship quality in engaged couples. In dating couples, neurotic characteristics were negatively correlated with relationship quality, whereas neuroticism was positively correlated with relationship quality in married couples [69]. In a political setting, a study found that individuals who score higher in neuroticism are more likely to have right-wing political affiliations [56]. As a result, it is apparent that the OCEAN personality model has applications in a wide variety of settings and can give a variety of insights about a person.

While the OCEAN personality model has had considerable impact in academia, the model has still been subject to a large amount of scrutiny from published studies. Some authors have used the terms ‘psychology of the stranger’ or a ‘cloudy measurement’ to describe the OCEAN model [94, 51]. The first of which is referring to the suggestions that the model only discovers traits that are easy to observe in a stranger [94]. Further to this, other authors have suggested the OCEAN personality inventory only accounts for 56% of the normal personality trait sphere (without considering the abnormal personality trait sphere) [21]. Another criticism of the OCEAN model is that it has limited scope; some researchers have suggested that it neglects other aspects of personality such as honesty, religiosity, thriftiness and others [112]. It has further been shown that the five factors in the model are not completely independent and rather measure human aspects which are inter-related [99]. Arguably the biggest criticism of the OCEAN personality model is that it's not based on any underlying psychological theory. It is merely an empirical finding which came from using factor analysis as a dimension reduction on a large dataset [17]. Block (2010) suggested the model uses factor analysis as an exclusive paradigm for conceptualising personality and reinforces that there is limited psychological theory which backs these findings [17, 21]. In Section 2.2.2 we discuss the Myers-Briggs Personality Model which is a theory driven model, rather than a statistically driven model.

2.2.2 Myers-Briggs Personality Model

The Myers-Briggs personality model was constructed by an American personality researcher, Katharine Cook Briggs and her daughter Isabel Briggs Myers [18]. It is recognised as the most well-known personality model due to its application to hiring processes, social dynamics, education and relationships [40, 156, 88]. However, the model has received a large amount of scrutiny, particularly from psychologists who question its validity and reliability [116, 63]. We discuss a number of these limitations in the latter parts of this section and provide evidence of why the model is useful in the context of social media research.

Briggs began researching personality in 1917 as she noticed clear differences in the personalities of individuals in her family [147]. As a result, Briggs and Myers both thoroughly studied the work of famous psychiatrist and psychoanalyst, Carl Jung. In particular, his English translated publication of 'Psychological Types' in 1923 [101]. Briggs and Myers endeavored to create practical use from the theory of personality types and consequently began formulating and testing the Myers-Briggs Type Indicator (MBTI) during the second world war [146]. Their aim was to identify the most comfortable and effective war-time jobs for women entering the industrial workforce for the first time [101]. The MBTI Handbook was then published in 1956 with second and third editions of the handbook being published in 1985 and 1998, respectively [100].

A majority of the MBTI handbook is based upon the theory proposed by Carl Jung; who speculated that humans experience the world through two dichotomous pairs of cognitive functions. Firstly, the rational/judging functions (thinking and feeling) and secondly the irrational/perceiving functions (sensation and intuition) [79, 73]. Jung proposed that humans use one of these four functions more primarily and dominantly than the other three; however, all four functions can be used at different times depending on the circumstances [101]. Moreover, Jung believed that each of these functions are expressed primarily in either an extroverted or introverted form [101]. Jung also suggested there was a conscious and unconscious combination of these functions [52]. In Jung's theory of psychological types, he used the terms dominant, auxiliary and inferior, where there is one dominant function, two auxiliary functions and one inferior function. The primary function is the most developed, differentiated and conscious function. Whereas the auxiliary functions are capable of more significant development or differentiation and help support the primary function [79]. The inferior function behaves more unconsciously and is always the opposite of the dominant function [101]. Additionally, Jung also introduces the concepts of a general attitude and a rationality. The general attitude categorises whether someone is extroverted/introverted and the rationality describes the preference for the primary functions, where thinking/feeling has a judging rationality and sensation/intuition has a perceiving rationality. Table 2.1 represents a summary of Jung's conception of the conscious personality types based on someones general attitude, rationality and their two dichotomous pairs of cognitive functions. Note that the unconscious combinations of the cognitive functions are omitted from Table 2.1 for brevity.

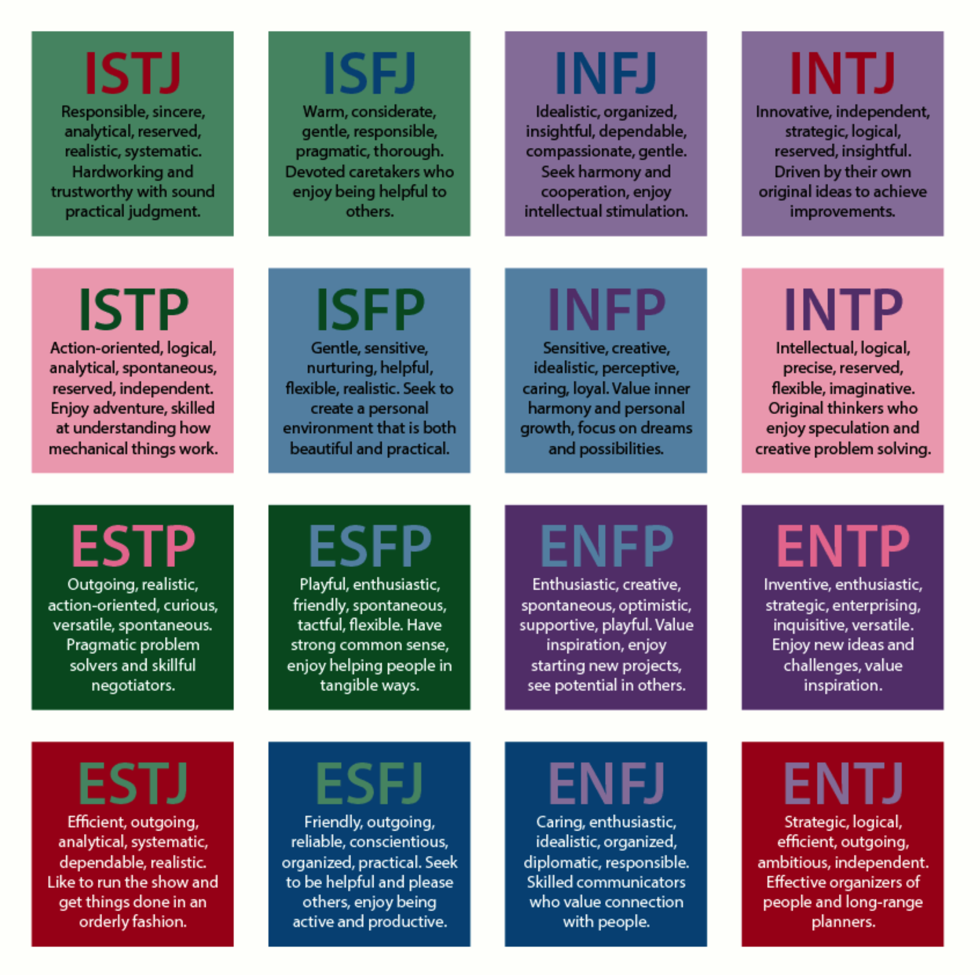
General Attitude	**Rationality**	**Primary**	**Auxiliary**		**Inferior**
Extroverted	Judging	Thinking	Sensation	Intuition	Feeling
		Thinking	Intuition	Sensation	Feeling
		Feeling	Sensation	Intuition	Thinking
		Feeling	Intuition	Sensation	Thinking
	Perceiving	Sensation	Thinking	Feeling	Intuition
		Sensation	Feeling	Thinking	Intuition
		Intuition	Thinking	Feeling	Sensation
		Intuition	Feeling	Thinking	Sensation
Introverted	Judging	Thinking	Sensation	Intuition	Feeling
		Thinking	Intuition	Sensation	Feeling
		Feeling	Sensation	Intuition	Thinking
		Feeling	Intuition	Sensation	Thinking
	Perceiving	Sensation	Thinking	Feeling	Intuition
		Sensation	Feeling	Thinking	Intuition
		Intuition	Thinking	Feeling	Sensation
		Intuition	Feeling	Thinking	Sensation

Table 2.1: Jungian model of conscious personality types.

The MBTI handbook utilises the general attitude function, the rationality function, the primary function and the first auxiliary function in Table 2.1 to form 16 unique personality types. The MBTI handbook represents these four cognitive functions through four dichotomous attitudes or functioning styles: Extroversion/Introversion (E/I), Intuitive/Sensory (N/S), Thinking/Feeling (T/F) and Judging/Perceiving (J/P). The first of which aims to describe how we interact with our environment; the preference to focus on the outer-world (extroverted) or your own inner-world (introverted). The second dichotomy aims to describe how people interpret information; the preference to interpret basic information and add meaning (intuitive) or to just focus on observed basic information (sensory). The third dichotomy aims to describe how people make decisions; to first consider logic and consistency (thinking) or to first consider people and special circumstances (feeling). Finally, the last dichotomy aims to describe structure; the preference to get things decided when dealing with the outside world (judging) or to stay open to new information and options (perceiving) [102]. As a result, the MBTI handbook illustrates a four factor model of personality where people attain one attribute from each of the four dichotomies; forming their 'personality type'. This results in 16 different personality types where a letter from each dichotomy is taken to produce a four letter acronym such as 'ENTJ' or 'ISFP'. Each of the 16 different personality types are provided in Figure 2.2 where the same background colour denotes mutual primary functions, the same text for the four letter acronyms denotes mutual auxiliary functions and the same black/white text denotes mutual general attitude functions. When designing the MBTI handbook, Briggs and Myers aimed to address two related goals: the identification of basic preferences of each of the four dichotomies which are implicit in Jung's theory and the identification and description of the 16 distinctive personality types that result from the interactions among the preferences [145].

Formulating a relevant and robust personality test was a large and important aspect of the MBTI handbook. Briggs and Myers designed an introspective self-report questionnaire where the responses to the questionnaire are used to determine someones personality type [101, 145, 103]. The questions which form the basis of the MBTI can be found in [146]. There are several websites providing online Myers-Briggs personality tests including `www.16personalities.com`, `www.mbtionline.com` and `www.truity.com`. It is also possible to find a certified MBTI professional to administer the test for you via the official Myers-Briggs website: `www.myersbriggs.org`. When undertaking the test, participants earn point scores for each dichotomy based on their responses to the questionnaire and the leaning of their final score determines their preference for that dichotomy. For instance, a user who scores higher for extroverted compared with introverted would be classified as an extrovert. Some websites providing a Myers-Briggs personality test also add a degree to how strong the preference is for each dichotomy. However, Briggs and Myers always considered the direction of the preference to be more important than the degree/strength of the preference. This is because a higher degree does not necessarily mean someone displays stronger characteristics for that attribute; rather they simply have a clearer preference for it [123].

While Briggs and Myers did utilise much of Jung's theory, there are some clear differences between the MBTI and the Jungian model. The most notable difference from



ISTJ
Responsible, sincere, analytical, reserved, realistic, systematic. Hardworking and trustworthy with sound practical judgment.

ISFJ
Warm, considerate, gentle, responsible, pragmatic, thorough. Devoted caretakers who enjoy being helpful to others.

INFJ
Idealistic, organized, insightful, dependable, compassionate, gentle. Seek harmony and cooperation, enjoy intellectual stimulation.

INTJ
Innovative, independent, strategic, logical, reserved, insightful. Driven by their own original ideas to achieve improvements.

ISTP
Action-oriented, logical, analytical, spontaneous, reserved, independent. Enjoy adventure, skilled at understanding how mechanical things work.

ISFP
Gentle, sensitive, nurturing, helpful, flexible, realistic. Seek to create a personal environment that is both beautiful and practical.

INFP
Sensitive, creative, idealistic, perceptive, caring, loyal. Value inner harmony and personal growth, focus on dreams and possibilities.

INTP
Intellectual, logical, precise, reserved, flexible, imaginative. Original thinkers who enjoy speculation and creative problem solving.

ESTP
Outgoing, realistic, action-oriented, curious, versatile, spontaneous. Pragmatic problem solvers and skillful negotiators.

ESFP
Playful, enthusiastic, friendly, spontaneous, tactful, flexible. Have strong common sense, enjoy helping people in tangible ways.

ENFP
Enthusiastic, creative, spontaneous, optimistic, supportive, playful. Value inspiration, enjoy starting new projects, see potential in others.

ENTP
Inventive, enthusiastic, strategic, enterprising, inquisitive, versatile. Enjoy new ideas and challenges, value inspiration.

ESTJ
Efficient, outgoing, analytical, systematic, dependable, realistic. Like to run the show and get things done in an orderly fashion.

ESFJ
Friendly, outgoing, reliable, conscientious, organized, practical. Seek to be helpful and please others, enjoy being active and productive.

ENFJ
Caring, enthusiastic, idealistic, organized, diplomatic, responsible. Skilled communicators who value connection with people.

ENTJ
Strategic, logical, efficient, outgoing, ambitious, independent. Effective organizers of people and long-range planners.

Figure 2.2: Personality types in the Myers-Briggs Type Indicator [128].

Jung's original thoughts is the concept that people only exhibit the conscious personality types, reducing the number of unique personality types from 32 to 16. Some researchers suggest that admitting the unconscious personality types is careless and "hardly fair to Jung" [52]. However, Briggs and Myers suggest that everyone has a conscious personality type, with some being more conscious than others. It is also apparent that the latter part of Jung's theory did not involve questionnaire measurement [25]. Moreover, Briggs and Myers introduce a fourth aspect to their model which is not present in the Jungian model: that people also have a preference for using either the judging function or the perceiving function. According to Myers, the judging/perceiving dimension aims to capitulate peoples preference to "have matters settled" (judging) or consider new options (perceiving) [101]. Another clear distinction from Jung's theory is that Briggs and Myers saw the dichotomies as dualistic: people have a clear preference for each category. Whereas, Jung saw the dichotomies as tendencies: humans have both and people can be balanced [10, 79]. This is consistent with why Jung's theory was surrounding personality types and not personality tests; Jung's theory was not designed to be applied in the ways it has now [80]. In any case, it is important to note that both models remain hypothetical, with no controlled scientific studies supporting Jung's original concept of personality types or the Myers-Briggs variation of this [25].

Many psychologists have scrutinised personality tests since their more frequent usage in the hiring processes of a number of workplaces. These practitioners question whether personality tests 'really capture who you are' and there is growing suspicions that humans may be growing tendencies to enjoy dividing people into categories [159, 13]. Nonetheless, it is apparent that the use of self-report questionnaires rest on several assumptions. The underlying assumption of the MBTI is that we all have specific preferences in the way we construe our experiences, and these preferences underlie our interests, needs, values, and motivation [115]. We also assume that there are general cognitive systems which can be expressed in terms of numerical scores on a measured scale. We assume these systems can be quantitatively evaluated by totaling the responses which lie in the same personality dimension. If there is a correlation between someones response on a questionnaire and that personality dimension, then we assume these attributions must be linked; introducing the potential of confirmation bias [153]. Smit (1983) proposed a number of requirements for these self-report questionnaires to be practical. These requirements are: the test must be comprehensive without being too time consuming, the test needs to be standardized for a specific population, test instructions should be clear, and the psychometric qualities of the test should yield reliable and valid results [122].

The MBTI instrument's validity has been subject to much criticism since its publication in 1956. Some sources have called the test 'pretty much meaningless' and the 'fad that won't die' [91, 63]. One primary concern of many researchers is the very little evidence for the dichotomies. As it was previously mentioned, Briggs and Myers always considered the direction of the preference to be more important than the degree/strength of the preference [123]. Hence, scores on each of the four MBTI dimensions should follow a bimodal distribution with a higher proportion of people scoring towards the end of each tail. However, a majority of studies have shown that the distribution of scores for each dimension is fairly bell shaped [14]. This indicates that a majority of people portray more balanced characteristics for each of the four dimensions, a result consistent with Jung's theory: that the dichotomies are more like tendencies: humans have both and people can be balanced [10, 79]. Hence, researchers believe it may be more appropriate to report scales along each of the four distributions rather than using a cutoff value to form the dichotomies [116, 95].

Other research has found that the validity and utility of the MBTI is problematic. In 1991, The United States National Academy of Sciences committee reviewed data from the

MBTI and found that on the extrovert/introvert dichotomy had high correlations with comparable scales of other personality inventories. Moreover, the intuitive/sensory and thinking/feeling dichotomies showed weak correlations with comparable scales and as a result, the committee concluded there was "not sufficient, well-designed research to justify the use of the MBTI in career counseling programs". Note that the study formulated its measurement of validity based on "whether the MBTI predicts specific outcomes related to interpersonal relations or career success/job performance" [107].

The accuracy of the MBTI fundamentally depends on the honesty of participants self-reporting. While this may seem like an issue in most self-reporting personality questionnaires, it is possible to reduce these effects by utilising validity scales to assess if exaggerated or socially desirable responses are present [7]. This also significantly reduces the likelihood of confirmation bias in the results. Personality tests such as the 16PF Questionnaire and the Minnesota Multiphasic Personality Inventory make use of these validity scales. However, the MBTI does not use validity scales and as a result, fundamentally relies on honest self-reporting. One study used the Eysenck Personality Questionnaire lie scale to determine how socially desirable people were being in their responses to the questionnaire. The study found there to be a weak correlation between the judging/perceiving dichotomy and the lie scale; indicating a small proportion of people are likely to be lying in their responses relating to the judging/perceiving category [54].

The reliability of the MBTI tends to be questioned by some researchers. According to Briggs and Myers, your personality type is inborn and doesn't change. However, a considerable number of people have obtained different personality types when retaking the questionnaire within months [116, 63]. One study found that about 50% of people obtain the same overall type when retested within 9 months and around 36% of people obtain the same overall type when retested after 9 months [24]. As a result, researchers have questioned whether Briggs and Myers were correct in postulating that a personality type is something people are born with. Really this condenses to the age old debate of nature/nurture and research has shown that it is a combination of both factors that determine someone's personality [44]. Jung theorised that people have an innate urge to grow and "psychological types is the compass guiding this growth process" [145]. The general consensus among researchers in this area suggests that as people grow, their thinking changes and they most likely develop within their type; but it is unlikely for their underlying type to change [81, 19]. Another study surveyed people on what their preferred personality type was compared to that assigned by the MBTI and only half of the people chose the same type [26]. As a result, this study presents convincing evidence that individuals don't experience the Barnom effect when sitting the MBTI [25].

While many researchers are divided over the usefulness of the MBTI, it is important to remember that human personalities are very complex; personality types are just a way we represent these highly dimensional phenomena in a low dimensional and easily understood setting. As with any lower dimensionality representation, it is subject to a higher degree of error compared to its higher dimensional counterpart. Consequently, it is no surprise that personality tests receive some criticism by researchers. However, personality tests certainly aren't completely useless – there is increasingly more research which suggests that personality tests can provide useful information about yourself [57]. Their practicality has been recognized in the areas of work, social dynamics, education and relationships [40, 156, 88]. One large-scale study showed there were clear connections between personality traits and occupations; suggesting some personality inventories are appropriate for use in a professional context [133]. Another study found that the MBTI is useful in helping employees and supervisors in human service organizations become more aware of conflicts and agreements in the workplace [8]. Moreover, there has been a great amount of research investigating the impact of Myers-Briggs personality types on team effectiveness at universities and other educational environments. One study conducted on engineering students found that communication, trust and interdependence were improved when individuals were trained on the personality type of students in their team [151]. Another study on business students found that macroeconomic students reported greater satisfaction when students were working with similar personality types to their own. Whereas, the same study found that marketing students preferred to work with personality types dissimilar to their own [4]. Other studies have suggested there are certain personality types which make a romantic relationship more compatible and satisfying. A study on the relationship between marital satisfaction and the MBTI found that the extrovert/introvert dichotomy effected a couples' satisfaction on the scales of time spent together and affective communication; where two married introverts suffered mostly in these areas [67]. The same study found that two married introverts were also more likely to be maritally satisfied [67]. From all of these findings, it is apparent that the MBTI has a wide variety of applications in settings that can give a number of meaningful insights about a person.

Clearly the Myers-Briggs personality model has been strongly criticised by academics, particularly those in psychology. However, there is a number of reasons why the model is still useful and can be appropriately applied to personality profiling on social media. We utilise the Myers-Briggs model as the primary model in our analysis because of the following reasons:

- Thousands of Twitter users self-report their MBTI on Twitter. This enables us to avoid long, cumbersome and expensive questionnaires, and instead obtain a labelled dataset through appropriately querying for each of the 4 letter personality type acronyms.

- The Myers-Briggs acronyms are unique to Myers-Briggs (to our knowledge) and thus won't be confused with any other acronym, reducing the potential of any incorrectly labelled users.

- The Myers-Briggs model is recognised as the most well-known personality model and thus has the largest number of self-reports on Twitter, enabling us to achieve the largest labelled personality dataset on Twitter.

- The Myers-Briggs model is relatively under-studied on social networks compared to the OCEAN model, meaning there is the potential for more substantial advances in the literature surrounding the Myers-Briggs model.

- Finally, the aim of this work is to make a methodological contribution and develop a framework for modelling personality profiles from social media data using statistical ML approaches. The MBTI is is just used as a test case because of the reasons above, but we expect that the framework can be applied to other personality models (or other labelings/characteristics of individuals on social media) more generally.

2.3 Natural Language Processing Tools

In this Section, we introduce a number of Natural Language Processing (NLP) tools which will be utilised throughout our results.

2.3.1 Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC; pronounced "Luke") is a library of dictionaries which aim to capture people's social and psychological states based on the language they use. The foundations of LIWC were developed by American social psychologist, James Pennebaker. Pennebaker's key research findings established how human usage of 'low-level words', such as pronouns and function words, can be indicative of their large-scale behaviors, [157, 48].

Each dictionary comes with a list of words, word stems, emoticons and other specific verbal constructs which have been identified to portray a psychological category of interest. Note that we use the LIWC 2015 English dictionary in this paper. This dictionary is comprised of almost 6,400 words which follow a one-to-many mapping to 74 different word categories. These word categories include items such as swear words, function words, past focused words, negative emotion words and more. A full description of these word categories, along with examples is provided in the paper by Pennebaker *et al.* [113].

When using LIWC, we provide it already tokenized text. LIWC then compares each word in the text to its dictionary of words and calculates the proportion of total words in the text that match each of the 74 different categories.

2.3.2 Valence Aware Dictionary and Sentiment Reasoner

Valence Aware Dictionary for Sentiment Reasoning (VADER) is a simple rule-based model for sentiment analysis. VADER is constructed based upon a gold-standard list of lexical features and their associated sentiment intensity measures which are attuned to sentiment in microblog-like contexts [74]. The list of lexical features includes English words, emoticons and slang. These features were created by collecting ratings from humans on Amazon Mechanical Turk. Lexical features were rated on a scale from '-4' (extremely negative) to '+4' (extremely positive), with '0' representing neutral (or NA). The mean rating for each lexical term was then used to represent its sentiment polarity (positive/negative) and its sentiment intensity (on a scale from -4 to +4). These lexical features are then combined with five general rules that aim to embody grammatical and syntactical conventions for expressing and emphasising sentiment intensity. These incorporate changes in sentiment polarity and intensity due to punctuation, capitalisation, degree modifiers, contrastive conjunctions and negation. These rules are discussed more thoroughly by Hutto and Gilbert [74].

For any given text, VADER will return four scores, referred to as: positive, negative, neutral and compound scores. The positive, negative and neutral scores represent the proportion of positivity, negativity and neutrality in a given string, respectively. Whereas, the compound score represents a normalisation of the total sentiment and is represented by a number between -1 and +1. The normalisation used is:

$$\frac{\phi}{\sqrt{\phi^2 + \alpha}},$$

where ϕ is the sum of the mean sentiment scores for all lexical features and α is a normalisation parameter that is set to 15 [23]. Thus, the compound score represents a normalisation of the total sentiment, where scores closer to +1 indicate more positivity, scores closer to -1 indicate more negativity and scores around zero indicate neutrality.

2.3.3 Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique which is used for natural language processing. BERT was created and published in 2018 by Google employee, Jacob Devlin and his colleagues [41]. Now, Google utilises BERT in almost every English-language query to its search engine and a number of authors have suggested that "BERT has become a ubiquitous baseline in NLP experiments" [130]. BERT works by tokenizing natural language and creating 768-dimensional output embedding vectors for each of the tokens which aim to portray their meaning. Words which are closer together in the 768-dimensional vector space tend to have a more similar meanings and words further apart tend to have differing meanings. It is believed that the BERT dimensions all correspond to different representations of linguistic and semantic meaning, but this has been difficult to verify rigorously [32].

The BERT model was pretrained on two tasks: language modelling (where 15% of tokens were masked and BERT was trained to predict their embeddings from the context) and next sentence prediction (where BERT was trained to predict the next sentence in a given document). After pretraining, BERT can be finetuned on downstream tasks with significantly less resources. The BERT model produces contextualised embeddings, a recently developed technology which doesn't fix the embeddings vector of a given token. Other embedding models like Word2vec and GloVe have fixed embeddings vectors for every token. The reason contextualised embeddings are important is because words often have different semantic meanings. For instance, the token 'bank' has a number of different meanings including; the land alongside a river/lake, a high mass/mound of a particular substance or a financial establishment that uses money. Contextually dependent models such as BERT will aim to represent the embeddings of such words differently, depending on their meaning and context.

The original English-language BERT was trained on data extracted from BooksCorpus (with 800M words) and English Wikipedia (with 2,500M words). The original BERT has two models: BERT based (with 110M parameters) and BERT large (with 345M parameters). However since the original model was published in 2018, scientists have created a number of different BERT models which are trained on either different languages or different contexts. One example of this is the BERTweet model, which is the BERT model we used for all analysis in this paper [105]. BERTweet has the same model architecture as the original BERT models and is the first public large-scale language model pre-trained on English Tweets. The corpus used to pre-train the model consisted of 850M English Tweets, containing 845M Tweets streamed from January 2012 until August 2019 and 5M Tweets related to the Covid-19 pandemic [105].

2.3.4 Botometer

Botometer is an Application Programming Interface (API) based upon a supervised machine learning classifier that distinguishes bot-like and human-like accounts through utilising account and language features. The classifier is fundamentally based upon 15 labelled datasets which include human-like and bot-like accounts [163]. The datasets are annotated using various methods including human annotation, observations of self declaration and signs of coordination. The authors of Botometer observed that different types of bots tend to have unique behavioral patterns [135]. As a result, Botometer groups the type of bots into six categories: Astroturf, Fake Follower, Financial, Self Declared, Spammer and Other. Independent random forest classifiers are then trained on over 1000 different features and used to model each type of bot as well as the human accounts [163]. The outputs of these models are used in an Ensemble of Specialised Classifiers (ESC) architecture to model the overall likelihood the account is a bot – Sayyadiharikandeh *et al.* [135] discuss this method in detail.

For every account, Botometer calculates the features based upon the most recent 200 tweets and tweets mentioning the account [163]. The features can be characterized into six different classes: user profile, content/language, network, temporal, sentiment and friends

[20]. For example, user profile characteristics include features such as the length of the screen name, the account location and the age of the account [163]. Content/language features include information such as the number of verbs, nouns or adjectives in the text [45]. Network features include information on the users' retweets, quote tweets, mentions and hashtags. Temporal features include attributes such as the tweet rate and timing patterns of retweets/quote tweets [163]. Sentiment features include information about the positivity/negativity of an accounts text (see Section 2.3.2) and friends' features include similar features discussed prior but regarding an accounts friends.

Botometer uses these features to return a 'probability' of an account being each type of bot class [20]. The authors additionally utilise the ESC architecture to provide a Complete Automation 'Probability' (CAP) that the account is a bot [135]. It is important to note that the content/language and sentiment features are only based on English language – so these features become irrelevant when a non-English account is passed to Botometer. Consequently, language-independent probabilities are provided for the previously mentioned bot types and CAP. Botometer additionally returns an accounts primary language in its API response and encourages subscribers can use these language-independent probabilities for non-English accounts [163].

2.4 Mathematical Background

2.4.1 Binary Models

In this section, we discuss a number of binary mathematical models. Suppose we have n samples of data for model training. We wish to model a binary response variable;

$$\{Y_i \ ; \ i = 1, \ldots, n\},$$

for a particular event given a set k features;

$$\{x_{i,j} \ ; \ i = 1, \ldots, n \ \& \ j = 1, \ldots, k\}.$$

The notation, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,k})$ is used to denote the set of features for sample, i. We let the response variable take a value of '1' if the event occurs and a value of '0' otherwise. We will use the notation $P(Y_i = 0)$ or $P(Y_i = 1)$ to denote the probability of observing of a particular outcome and $p(y_i) := P(Y_i = y_i)$ denote the probability of an arbitrary outcome – the same notation will be used for conditional probabilities and joint probabilities. In what follows, we describe a series of models for modelling the conditional probability;

$$p(y_i \mid \mathbf{x}_i) := P(Y_i = y_i \mid \mathbf{x}_i).$$

In each case, we determine the prediction, Y_0, for a given set of features, $\mathbf{x}_0$ by maximising this conditional probability;

$$Y_0 = \arg\max_y p(y_0 \mid \mathbf{x}_0) \quad y_0 \in \{0, 1\}.$$

Logistic Regression

Under the logistic regression model, we assume the log-odds of the response variable to be a linear combination of one or more features. That is;

$$\operatorname{logit}\left(P\left(Y_i=1 \mid \mathbf{x}_i\right)\right)=\log \left(\frac{P\left(Y_i=1 \mid \mathbf{x}_i\right)}{1-P\left(Y_i=1 \mid \mathbf{x}_i\right)}\right)=\beta_0+\sum_{j=1}^k \beta_j x_{ij},$$

where $\beta_0, \ldots, \beta_k$ represent a set of model parameters. The model is also often reformulated with respect to the event probability;

$$P\left(Y_i=1 \mid \mathbf{x}_i\right)=\frac{1}{1+\exp\left[-\left(\beta_0+\sum_{j=1}^k \beta_j x_{i,j}\right)\right]},$$

where the logistic function is defined as $f(x) := \frac{1}{1+\exp(-x)}$, hence where the logistic regression model got its name.

To estimate the parameters of the logistic regression, we perform maximum likelihood estimation. This is typically solved using numerical methods to maximise the log-likelihood;

$$\ell=\sum_{i=1}^n\left[Y_i \log \left(P\left(Y_i=1 \mid \mathbf{x}_i\right)\right)+\left(1-Y_i\right) \log \left(P\left(Y_i=0 \mid \mathbf{x}_i\right)\right)\right].$$

Naive Bayes

The naive Bayes classifier is a probabilistic classifier with strong independence assumptions between the predictors. Firstly, observe we can reformulate the conditional event probability using Bayes' Theorem;

$$p\left(y_i \mid \mathbf{x}_i\right)=\frac{p\left(\mathbf{x}_i \mid y_i\right) p\left(y_i\right)}{p\left(\mathbf{x}_i\right)} \propto p\left(\mathbf{x}_i \mid y_i\right) p\left(y_i\right).$$

Note that this is equivalent to modelling the joint probability;

$$p\left(y_i, \mathbf{x}_i\right),$$

which can be rewritten by repeatedly using the chain rule for probability;

$$\begin{aligned}
p\left(y_i, \mathbf{x}_i\right) &= p\left(x_{i,1}, \ldots, x_{i,k}, y_i\right) \\
&= p\left(x_{i,1} \mid x_{i,2}, \ldots, x_{i,k}, y_i\right) p\left(x_{i,2}, \ldots, x_{i,k}, y_i\right) \\
&= p\left(x_{i,1} \mid x_{i,2}, \ldots, x_{i,k}, y_i\right) p\left(x_{i,2} \mid x_{i,3}, \ldots, x_{i,k}, y_i\right) p\left(x_{i,3}, \ldots, x_{i,k}, y_i\right) \\
&= \ldots \\
&= p\left(x_{i,1} \mid x_{i,2}, \ldots, x_{i,k}, y_i\right) p\left(x_{i,2} \mid x_{i,3}, \ldots, x_{i,k}, y_i\right) \ldots p\left(x_{i,k} \mid y_i\right) p\left(y_i\right)
\end{aligned}$$

We then ‘naively’ assume that all the predictors are mutually independent, conditional on the response variable. That is;

$$p\left(x_{i,\ell} \mid x_{i,\ell+1}, \ldots, x_{i,k}, y_i\right) = p\left(x_{i,\ell} \mid y_i\right)\ \forall \ell.$$

where the joint probability then becomes;

$$p\left(y_i, \mathbf{x}_i\right) = p\left(y_i\right) \prod_{j=1}^{k} p\left(x_{i,j} \mid y_i\right).$$

Observe, that the conditional probability of the response variable becomes;

$$p\left(y_i \mid \mathbf{x}_i\right) = \frac{1}{C} \cdot p\left(y_i\right) \prod_{j=1}^{k} p\left(x_{i,j} \mid y_i\right),$$

where

$$C = p\left(\mathbf{x_i}\right) = P\left(Y_i = 0\right) p\left(\mathbf{x}_i \mid Y_i = 0\right) + P\left(Y_i = 1\right) p\left(\mathbf{x}_i \mid Y_i = 1\right)$$

is a scaling factor which is only dependent on the feature variables. We then estimate the class probabilities as;

$$P\left(Y_i = 0\right) = \frac{\sum_{i=1}^{n} I_{\{Y_i=0\}}}{n} \quad \& \quad P\left(Y_i = 1\right) = 1 - P\left(Y_i = 0\right).$$

Furthermore, we assume the likelihood of the features to have a Gaussian Distribution;

$$P\left(x_{i,j} \mid Y_i = \nu\right) = \frac{1}{\sqrt{2\pi\sigma_{j,\nu}^2}} \exp\left(-\frac{\left(x_{i,j} - \mu_{j,\nu}\right)^2}{2\sigma_{j,\nu}^2}\right),\ \nu \in \{0, 1\}$$

where $\mu_{j,\nu}$ and $\sigma_{j,\nu}^2$ is the mean and variance of the data $(x_{1,j}, \ldots, x_{n,j})$ associated with class ν.

Support Vector Machines

Support Vector Machines (SVMs) are a supervised learning model which constructs hyper-planes in high dimensional space. We use these hyper-planes to maximise the width of the gap between samples from each class. New examples are then mapped into the same space and predicted to belong to a category based on which side of the hyper-plane they lie on. Hence, SVMs work by finding a $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that the prediction given by;

$$\operatorname{sign}\left(w^T \phi\left(\mathbf{x}_i\right) + b\right)$$

is correct for most samples, where ϕ is some feature mapping. In order to solve this, we define the following primal problem;

$$\begin{array}{ll} \min\limits_{w,b,\zeta} & \frac{1}{2}w^T w + C\sum_{i=1}^{n}\zeta_i, \\ \text{subject to} & Y_i\left(w^T\phi\left(\mathbf{x}_i\right)+b\right) \geq 1-\zeta_i, \\ & \zeta_i \geq 0, \quad i=1,\ldots,n, \end{array}$$

where ζ_i is some distance/tolerance we allow each sample to be from the fitted hyperplane and C as an inverse regularization parameter which controls for the strength of the penalty associated with these distances. Intuitively, we are maximising the margin (by minimising $||w||^2 = w^T w$), while incurring a penalty for every misclassified sample which is within some allowable distance to the fitted hyperplane. This allowable distance from the hyperplane is called the marginal boundary and we define the support vectors to be any features within this permitted distance;

$$V = \left\{i\ ;\ 1-\zeta_i \leq Y_i\left(w^T\phi\left(\mathbf{x}_i\right)+b\right) \leq 1\right\}.$$

The dual problem to the primal is;

$$\begin{array}{ll} \min_\alpha & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha, \\ \text{subject to} & \mathbf{y}^T\alpha = 0, \\ & 0 \leq \alpha_i \leq C, \quad i=1,\ldots,n, \end{array}$$

where e is a vector of ones, $\mathbf{y} = (Y_1,\ldots,Y_n)$ is a vector of the outcomes and Q is an $n\times n$ positive semidefinite matrix with;

$$Q_{ij} = Y_i Y_j K\left(\mathbf{x}_i,\mathbf{x}_j\right),$$

where we define;

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right) = \phi\left(\mathbf{x}_i\right)^T\phi\left(\mathbf{x}_j\right)$$

as the Kernel. The terms α_i are called the dual coefficients and they are upper-bounded by C. This optimisation problem is then solved using gradient descent methods. Once the optimisation is complete, the decision function for a given set of features, $\mathbf{x}_0$ becomes;

$$f\left(\mathbf{x}_i\right) = \sum_{i\in V} Y_i\alpha_i K\left(\mathbf{x}_i,\mathbf{x}_0\right)+b,$$

and the predicted class corresponds to the sign of this decision function. Platt Scaling is used to formulate probability estimates for predictions [118]. Platt Scaling is an algorithm which transforms outputs from a classification model into probability estimates by utilising the logistic function;

$$p\left(y_i \mid \mathbf{x}_i\right) = \frac{1}{1+\exp\left(Af\left(\mathbf{x}_i\right)+B\right)},$$

where A and B are scalar parameters which are estimated using maximum likelihood methods.

Random Forests

Random forests are an ensemble learning method for classification tasks which constructs a multitude of decision trees during model training. The training algorithm for random forests applies a bootstrap aggregating technique, often referred to as bagging. Bagging is where we sample from our training data with replacement to generate a series of models and average the results from each model during prediction.

In the random forest algorithm, we apply bagging repeatedly (b times) to the training set and fit separate decision trees to each of these samples of data. A classification decision tree is a supervised machine learning algorithm where the feature variables are recursively split according to a loss function, constituting a tree-like structure. As a result, samples with the same labels or similar target values are grouped together. We refer to parts of the tree using graph terminology: nodes represent leaves and edges represent branches.

For a particular bootstrapped sample, $s \in \{1, \dots, b\}$, we initially draw n samples of data from the training set with replacement. Suppose we are at some node in the decision tree, m, and we wish to perform a candidate split. Let there be n_m samples of data at this node represented by;

$$Q_m \subset \{(x_{i,j}, Y_i) \mid i = 1, \dots, n \;\&\; j = 1, \dots, k\}.$$

For each candidate split $\theta = (\ell, t_m)$ consisting of feature ℓ and threshold t_m, we partition the data into subsets Q_m^{left} and Q_m^{right} with data n_m^{left} and n_m^{right}, respectively;

$$Q_m^{\text{left}}(\theta) = \{(x_{i,\ell}, Y_i) \in Q_m \mid x_{i,\ell} \leq t_m, i = 1, \dots, n\},$$

$$Q_m^{\text{right}} = Q_m \backslash Q_m^{\text{left}}(\theta).$$

We make use of the Gini Impurity loss function to determine the quality of a split at node m;

$$H(Q_m \mid \theta) = 2P(Y_i = 0 \mid \theta) P(Y_i = 1 \mid \theta),$$

where we determine the class probabilities for the split as following;

$$P(Y_i = y \mid \theta) = \frac{1}{n_m} \sum_{Y_i \in Q_m} \mathbb{1}_{\{Y_i = y\}}, \quad y \in \{0, 1\}.$$

The overall quality of a split at node m is then determined by calculating the overall impurity function;

$$G(Q_m \mid \theta) = \frac{n_m^{\text{left}}}{n_m} H\left(Q_m^{\text{left}} \mid \theta\right) + \frac{n_m^{\text{right}}}{n_m} H\left(Q_m^{\text{right}} \mid \theta\right),$$

and we select the parameters which minimise the overall impurity;

$$\theta^* = \arg\min_{\theta} G(Q_m \mid \theta).$$

This optimisation problem is solved using the CART algorithm, a greedy algorithm for selecting the most appropriate feature and its associated threshold [37]. We perform splits recursively for subsets $Q_m^{\text{left}}(\theta^*)$ and $Q_m^{\text{right}}(\theta^*)$ until the maximum allowable depth is reached. If no further splits are performed at a particular node, t, then we call this a terminal node and use the class probabilities at these nodes to represent the predicted class probability, $p_s(y_i \mid \mathbf{x}_i)$, of any data at this node, $(\mathbf{x}_i, Y_i) \in Q_t$.

A random forest is then fit using the results of the b decision trees on the training data. For a given set of features, $\mathbf{x}_i$ we follow a top-to-bottom path down each of the decision trees until we reach a terminal node. We then obtain the predicted class probabilities of the sample by averaging the predicted class probabilities from each of the decision trees;

$$p(y_i \mid \mathbf{x}_i) = \frac{1}{b} \sum_{s=1}^{b} p_s(y_i \mid \mathbf{x}_i).$$

2.4.2 Statistical Methods

In this section, we will present a number of different statistical methods including various statistics and hypothesis tests.

F-test

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. Here, we present the F-test such that it is formulated in a regression setting.

Suppose we have two regression models which are both fitted on the same dataset. Call these Model 1 and Model 2, where Model 1 has p_1 parameters and Model 2 has p_2 parameters. Further suppose that Model 1 is 'nested' within Model 2 such that $p_1 < p_2$. We call Model 1 the restricted model and Model 2 the unrestricted model.

The unrestricted model will always be able to fit the data at least as well as the restricted model, however we use an F-test to determine whether Model 2 gives a statistically significant better fit to the data. That is, we wish to test the null hypothesis;

H_0 : Model 2 does not provide a statistically significant better fit than Model 1.

The F-statistic for this hypothesis test is given by;

$$F = \frac{\left(\frac{\text{RSS}_1 - \text{RSS}_2}{p_2 - p_1}\right)}{\left(\frac{\text{RSS}_2}{n - p_2}\right)},$$

where RSS_1 and RSS_2 are the residual sum of squares of model 1 and model 2, respectively. Under the null hypothesis, F will have a F-distribution with $(p_2 - p_1, n - p_2)$ degrees of freedom.

Cramér's V

The Cramér's V Statistic is based on Pearson's chi-squared statistic and measures the association between two categorical variables, giving a value between 0 and +1 (inclusive), [36]. The Cramér's V Statistic is computed through utilising the Chi-Square Statistic and its associated p-value is returned from a Chi-Square Test.

Consider two categorical variables Y and Z of size n with r and k categories, respectively. Let $n_i = |Y_i|$, $n_j = |Z_j|$ and $n_{i,j}$ denote the number of times we observe the pair of values, (Y_i, Z_j). The Chi-Square Statistic is then defined to be

$$\chi^2 = \sum_{i,j} \frac{\left(n_{i,j} - \frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}},$$

and its p-value is calculated from the Chi-Square Distribution with $(r-1)(k-1)$ degrees of freedom. We then use a bias correction when calculating the Cramér's V Statistic since it is heavily biased and tends to overestimate the strength of the association. The bias corrected version of the Cramér's V Statistic is given by

$$\tilde{V} = \sqrt{\frac{\tilde{\varphi}^2}{\min\left(\tilde{k}-1, \tilde{r}-1\right)}},$$

where

$$\tilde{\varphi} = \max\left(0, \frac{\chi^2}{n} - \frac{(k-1)(r-1)}{n-1}\right) \quad \& \quad \tilde{k} = k - \frac{(k-1)^2}{n-1} \quad \& \quad \tilde{r} = r - \frac{(r-1)^2}{n-1}$$

The result of $\tilde{V}$ can be interpreted in the same way we would interpret a correlation, however we would need to look at the contingency table to understand the direction of the relationship, if it existed.

Granger Causality Test

The Granger causality test is used to determine whether one time series is useful in forecasting another time series. The test was developed by Clive Granger, who argued that causality in economics could be tested for by measuring the ability to predict the future values of a time series using prior values of another time series [62].

The test is defined by considering two stationary time series, $X = \{X_0, \ldots, X_n\}$ and $Y = \{Y_0, \ldots, Y_n\}$. Suppose we are interested in whether X Granger-causes Y over p time lags. We do this by fitting two linear models. The first model we include only the lagged values of Y:

$$Y_t = \alpha_{1,0} + \alpha_{1,1} Y_{t-1} + \cdots + \alpha_{1,p} Y_{t-p} + \epsilon_{1,t}, \tag{2.1}$$

where we define $\epsilon_{i,t}$ as the error term of model i at time t and $\alpha_{i,j}$ as the parameter of model i at lag j. Next, we augment the model to also include the lagged values of X:

$$Y_t = \alpha_{2,0} + \alpha_{2,1}Y_{t-1} + \cdots + \alpha_{2,p}Y_{t-p} + \beta_1 X_{t-1} + \cdots + \beta_p X_{t-p} + \epsilon_{2,t}. \tag{2.2}$$

The null hypothesis:

$$H_0 : X \text{ does not Granger-causes } Y,$$

is accepted if and only if no lagged values of X are retained in the regression model observed in Eq. 2.2. It is most common to test this Null Hypothesis through an F-test, however other variations of likelihood ratio tests can also be used.

2.4.3 Data Manipulation Methods

Principal Component Analysis

Principal Component Analysis (PCA) is a data manipulation technique which finds a low-dimensional representation of a set of features such that it contains as much variation in the data as possible. These lower dimensional representations are called principal components.

Suppose we have n subjects and k features in our data. Suppose we have standardized our data, X and the standardized features are represented as $X_1, X_2, \dots, X_k$. The first principal component of the data is the normalised linear combination of the predictors:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

that has the largest variance, where we refer to the elements $\phi_{11}, \phi_{21}, \dots, \phi_{p1}$ as the loadings. The second principal component is then the normalised linear combination of the predictors with the second highest variance, and so on. To find these components, we calculate the eigenvalues and eigenvectors of Σ:

$$\Sigma\boldsymbol{\alpha}_i = \lambda_i\boldsymbol{\alpha}_i, \quad i = 1, \dots, k.$$

where $\lambda_1, \dots, \lambda_k$ is the set of eigenvalues and $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_k$ is the corresponding set of eigenvectors. Suppose the j^{th} largest eigenvalue is λ_j and its corresponding eigenvector is $\boldsymbol{\alpha}_j$. The j^{th} principal component of X is then given by:

$$\boldsymbol{\alpha}_j^T X \quad j = 1, \dots, k.$$

The eigenvalues are then often used to determine the proportion of variance in the data that it is explained by each of the principal components. For instance, the proportion of variance of the data that is explained by the j^{th} principal component is given by:

$$\frac{\lambda_j}{\left(\sum_{i=1}^{k} \lambda_i\right)}.$$

SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is a commonly used upsampling method to solve problems with class imbalance in mathematical models. It does this by synthesising new minority class instances to balance the distribution of classes. This technique can be applied to increase the number of samples in a particular class to any desirable amount.

Suppose we have n subjects and k features in our data:

$$X = \{x_{i,j} \ ; \ i = 1, \dots, n \ \& \ j = 1, \dots, k\},$$

and the set of features for a given subject is denoted by: $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,k})$. Furthermore, suppose that some of these features are part of some minority class set, A. We generate samples in A by firstly finding the k-nearest neighbours of each $\mathbf{x}_i \in A$, where k is a hyper-parameter which is normally set to $k = 5$. Let these k-nearest neighbours of $\mathbf{x}_i$ be denoted by A_i.

We then iterate over $\mathbf{x}_i \in A$ and $\mathbf{x} \in A_i$ to linearly interpolate new data in the minority class;

$$\mathbf{x}^{\text{new}} = \mathbf{x}_i + \lambda \cdot (\mathbf{x} - \mathbf{x}_i), \quad \lambda \sim U(0,1),$$

until we have synthesised the desirable amount of data.

2.5 Literature Review

In this section, we present a literature review of the work which is to be presented in the subsequent chapters.

2.5.1 OCEAN Personality Model

We provide an overview of the OCEAN Personality Model in Section 2.2.1 including its psychological framework and the development/history of the model. In this section, we will summarise the literature to date which has utilised the OCEAN model in social network research, with a strong focus on modelling OCEAN characteristics.

The OCEAN personality model has been the primarily used personality model for understanding the personal characteristics of users on social media. This is because of its popularity among academics, particularly data scientists, who value the model because it was created using a statistically driven approach [17]. Another reason for its popularity among data scientists is the easily attainable datasets which have existed over the past couple of decades. The first and most popular dataset was created in 2007 by an organisation called ‘myPersonality’. The organisation utilised a Facebook App which allowed its users to participate in psychological research by filling in a personality questionnaire [87, 137]. This allowed the authors to measure the OCEAN personality type of respondents and represent each of the 5 corresponding traits on a numerical scale (from low to high preference for the trait). The application was active until 2012 and collected data from over 6 million volunteers during this time. Around 40% of the respondents also opted to share their data from their Facebook profile, resulting in a dataset of Facebook features and personality features [84]. The authors provided other scholars with anonymized data to be used for non-commercial academic research resulting in dozens of peer-reviewed papers.

Much of the research utilised the myPersonality data to understand how certain Facebook characteristics shaped someones OCEAN personality type. For instance, Howlader *et al.* [71] explored the use of regression models for predicting personality traits. The authors used a basic Latent Dirichlet Allocation (LDA) with up to four topics as the features and compared the mean square error of the models with and without the inclusion of LIWC features. They observed that incorporating LIWC features significantly increased performance of the regression models.

Most other academics are using more of a machine learning approach to solving the problem [137, 143, 12, 144, 1]. Schwartz *et al.* [137] were the first set of authors to utilise this dataset. These authors performed a correlation analysis of peoples LIWC scores with gender, age and the OCEAN personality traits. They then used word clouds to represent the phrases which are most distinguishing for each of the traits. Finally, they utilised Ridge Regression to model the personality types of users from point-wise mutual information (PMI) scores, LDA topics and LIWC features based on their textual features. The authors utilised the square root of the coefficient of determination (R) as their evaluation metric and reported the best results when considering all features, with an average R value of 0.35.

Tadesse *et al.* [143] analyse and compare four machine learning models in predicting personality traits and calculate correlations between each of the feature sets and personality traits. Namely, they compare the accuracy of logistic regression, SVM, gradient boosting and XGBoost. The authors utilise the LIWC and SPLICE dictionaries for feature extraction as well as various social network analysis (SNA) features. They compare the results with and without each feature set and find that the XGBoost classifier outperforms other classifiers for all variations of the three feature sets, with an average prediction accuracy of 66.38%. Note that the highest prediction accuracies were achieved using an XGBoost model on strictly the SNA features, achieving accuracies of 73.3%, 68.0%, 65.3%, 69.8%, 78.6% for the Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism categories, respectively.

Başaran and Ejimogu [12] used a slightly different approach and modelled OCEAN personality types using a Recurrent Neural Network (RNN) with 5 output layers (one for each trait). The authors use 18 different features based on users personal information (age, gender, relationship status) and their facebook activity (events attended, groups joined, friends) to achieve an overall accuracy of 85% on the testing data. While these authors achieved a very high prediction accuracy, their inclusion/exclusion criteria was very stringent due to features like events attended and groups joined not existing for a lot of users. Furthermore, a lot of these features are not easily generalisable to other social media platforms like Twitter and Reddit where groups and events don't exist in the same way they do on Facebook. It is also important to note the presence of substantial class imbalance in the myPersonality dataset, particularly the dataset used by Başaran and Ejimogu [12], where some classes are up to 28 times larger than their counterpart. As a result, it is important to make the distinguishment between how informative a model is on the data and a models ability to achieve high accuracy on the data. We will refer to the prior as model performance and the latter as model prediction. Neither Tadesse *et al.* [143] or Başaran and Ejimogu [12] appear to address the class imbalance problem when using the myPersonality dataset. While their reported accuracies give clear evidence of high model prediction, these may not be representative of model performance.

Tandera *et al.* [144] acknowledge the class imbalance problem and address it by either down-sampling the majority class or up-sampling the minority class. The authors use LIWC, SPLICE and SNA features in a number of different machine learning and deep learning models. The machine learning models considered are: naive Bayes, SVM, logistic regression, gradient boosting and LDA. The deep learning models considered are: Multi-Layer Perceptron (MLP), Long Short Term Memory (LSTM), Gated Recurrent Units (GRU) and Convolutional Neural Network (CNN). The deep learning models outperformed the machine learning models for each trait. On average, MLP (with down-sampling) performed the best on the myPersonality data with an average accuracy of 70.78%.

Alam *et al.* [1] also acknowledges the presence of class imbalance in the myPersonality dataset. The authors address this issue by considering a macro-averaged variation of precision, recall and the F1 score when considering model performance. As a result, these metrics give clearer evidence of model performance over model prediction. However, the underlying model is still likely to significantly favor the majority class when performing predictions on unseen data. The authors use a bag-of-words approach to creating features and create unigrams for each of the tokens. They then fit a number of different models including a SVM, a Bayesian logistic regression and a Multinomial Naive Bayes (MNB) sparse model. The MNB model achieved the highest average accuracy of 61.79%.

Cambridge Analytica (CA) was a British political consulting firm who made use of the OCEAN personality model for what it described as 'behavioral micro-targeting' to predict the 'needs' of individuals and how these may change over time. Cambridge Analytica was formed in 2013 and was accused of being a global election management agency, leading the company to subsequently close operations in 2018. The company primarily utilised Facebook data and captured most of its data through online surveys which were conducted on an ongoing basis [162]. The data scientists behind the company built a multi-step co-occurrence type regression model, similar to Singular Value Decomposition

(SVD) and other matrix factorisation methods. Dimensionality reduction of the Facebook data was said to be the core of the model [68]. Sumpter [142] analysed the accuracy of CA's regression models in his book, 'Outnumbered'. He used a publicly available dataset of around 20,000 anonymized Facebook Users which was produced by Michael Kosinski, a psychologist who worked with CA. He found that the regression models used by CA could correctly classify the Openness category with 67% accuracy and each of the Conscientiousness, Extroversion, Agreeableness and Neuroticism categories with around 60% accuracy. As a result, Sumpter [142] suggested that personality assessment from online digital footprints does not work well. However, CA captured significantly more training data than him and he noted that CA's models contained much more personal information about the users, such as their geolocation, health data and census data.

Other authors have looked at predicting the OCEAN personality type of users on social media platforms other than Facebook, including Twitter [60, 125]. Golbeck *et al.* [60] created a Twitter application with two functions: to administer a 45-question version of the OCEAN personality test and to collect the most recent 2,000 tweets from each user undertaking the questionnaire. From this data, the authors collected features on these accounts including LIWC, Machine Reading Comprehension (MRC), sentiment and twitter based features. They fit two different regression algorithms: Gaussian Process and ZeroR. The results of the two different algorithms are then compared using the mean absolute error (MAE) on a normalised scale. They observe the GaussianProcess algorithm performs better for the Openness, Conscientious and Neuroticism categories, while the ZeroR performs better on the Agreeableness and Extroversion categories.

Quercia *et al.* [125] also models the OCEAN personality type of Twitter accounts. However, the authors collect data through obtaining the myPersonality dataset and considering the users who specified their Twitter accounts on their Facebook profiles, resulting in 335 Twitter accounts. The authors considered the correlation between Twitter features and the five OCEAN personality traits. They find that users with high neuroticism have low followers and following counts, whereas extroverted users tend to have high followers and following counts. The authors then predict the personalities of users based on only three inputs: number of people following, number of followers and their listed count. They use a regression model and report an average Root Mean Square Error (RMSE) of 0.79.

Table 2.2 provides an overview of the literature modelling the OCEAN personality type of users on Facebook and Twitter. Since authors provide different metrics when assessing a models performance and some model the traits altogether, it is difficult to evaluate which personality model is performing best. From Table 2.2, we observe that the RNN model by Başaran and Ejimogu [12] produces the largest accuracy of 85% for predicting the correct 'overall' personality type. This accuracy far exceeds any of the other accuracies and is also a multi-classification accuracy, making it even more impressive. However, it is highly likely that this model is predicting a significant majority of one personality type due to this dataset being very unbalanced and the authors providing no evidence of using any class imbalance techniques. As a result, we also expect the models performance to be fairly low. The second most accurate models then depends on the trait being measured – the models by Tadesse *et al.* [143] perform best on the Conscientiousness and Agreeableness traits, whereas the models by Tandera *et al.* [144] perform best on the Openness, Extroversion and Neuroticism traits. Since Tandera *et al.* [144] also address the class imbalance problem in their models, their accuracy metric is also likely a decent indicator for model performance. From these results, we also observe that Openness is consistently the easiest personality trait to model, achieving the best scores for the chosen metric in every authors results. Extroversion and neuroticism seem to be the most difficult personality traits to model, especially when using Twitter data. The CA models also perform significantly worse than the machine learning and deep learning models by Tadesse *et al.* [143], Tandera *et al.* [144]. This is likely because Sumpter [142] was limited in the data he could collect and that CA's was able to capture a lot more personal information about a user.

Authors	Data	Features	Model	Traits	Metric	Value
[71]	FB	LDA,LIWC	regression	O	MSE	0.43
[71]	FB	LDA,LIWC	regression	C	MSE	0.52
[71]	FB	LDA,LIWC	regression	E	MSE	0.64
[71]	FB	LDA,LIWC	regression	A	MSE	0.49
[71]	FB	LDA,LIWC	regression	N	MSE	0.50
[137]	FB	PMI,LDA,LIWC	regression	O	R	0.42
[137]	FB	PMI,LDA,LIWC	regression	C	R	0.35
[137]	FB	PMI,LDA,LIWC	regression	E	R	0.38
[137]	FB	PMI,LDA,LIWC	regression	A	R	0.31
[137]	FB	PMI,LDA,LIWC	regression	N	R	0.31
[143]	FB	SNA	XGBoost	O	Accuracy	73.3
[143]	FB	SNA	XGBoost	C	Accuracy	68.0
[143]	FB	SNA	XGBoost	E	Accuracy	65.3
[143]	FB	SNA	XGBoost	A	Accuracy	69.8
[143]	FB	SNA	XGBoost	N	Accuracy	78.6
[12]	FB	FB Specific Features	RNN	All	Accuracy	85.0
[144]	FB	LIWC,SPLICE,SNA	MLP	O	Accuracy	79.3
[144]	FB	LIWC,SPLICE,SNA	GRU	C	Accuracy	62.0
[144]	FB	LIWC,SPLICE,SNA	MLP	E	Accuracy	79.0
[144]	FB	LIWC,SPLICE,SNA	CNN	A	Accuracy	67.4
[144]	FB	LIWC,SPLICE,SNA	MLP	N	Accuracy	79.5
[1]	FB	bag-of-words	MNB	O	Accuracy	69.5
[1]	FB	bag-of-words	MNB	C	Accuracy	59.3
[1]	FB	bag-of-words	MNB	E	Accuracy	58.6
[1]	FB	bag-of-words	MNB	A	Accuracy	50.2

[1]	FB	bag-of-words	MNB	N	Accuracy	62.4
[142]	FB	FB Specific Features	CA Model	O	Accuracy	67
[142]	FB	FB Specific Features	CA Model	C	Accuracy	60
[142]	FB	FB Specific Features	CA Model	E	Accuracy	60
[142]	FB	FB Specific Features	CA Model	A	Accuracy	60
[142]	FB	FB Specific Features	CA Model	N	Accuracy	60
[60]	Twitter	LIWC,MRC,Sentiment	regression	O	MAE	0.119
[60]	Twitter	LIWC,MRC,Sentiment	regression	C	MAE	0.146
[60]	Twitter	LIWC,MRC,Sentiment	regression	E	MAE	0.160
[60]	Twitter	LIWC,MRC,Sentiment	regression	A	MAE	0.130
[60]	Twitter	LIWC,MRC,Sentiment	regression	N	MAE	0.182
[125]	Twitter	3 Twitter Features	regression	O	RMSE	0.69
[125]	Twitter	3 Twitter Features	regression	C	RMSE	0.76
[125]	Twitter	3 Twitter Features	regression	E	RMSE	0.88
[125]	Twitter	3 Twitter Features	regression	A	RMSE	0.79
[125]	Twitter	3 Twitter Features	regression	N	RMSE	0.85

Table 2.2: An overview of the literature to date which has modelled the OCEAN personality type of social media users. We present the best models from each paper, as determined by the metric chosen by each of the authors. Most authors model each trait independently, however Başaran and Ejimogu [12] model the traits altogether and do not assume independence. While it appears that Başaran and Ejimogu [12] achieve the best accuracy, it is important to note that the dataset used in their analysis was incredibly unbalanced and the authors appear to make no attempt to deal with class imbalance in the model. Hence, it is likely that the model performance is low for this model. Note that some models have formulated a classification problem and others a regression problem, resulting in the use of different metrics. The former assumes someone either displays the trait or they do not and the latter represents the trait on a scale.

2.5.2 Myers-Briggs Personality Model

In Section 2.2.2 we provided an overview of the Myers-Briggs personality model, including is development, history and psychological framework. In this section, we will summarise the literature to date which has utilised the Myers-Briggs personality model in social network research, with a main focus on modelling Myers-Briggs characteristics.

The Myers-Briggs personality model has been subject to very limited research in the social network domain when compared to the OCEAN model. This is likely because of the OCEAN models popularity among academics, but also because open-source labelled training data with Myers-Briggs personality types has not existed until recently. In 2017, Jolly [76] posted a labelled dataset on Kaggle, an online community of data science and machine learning practitioners. The dataset was comprised of 8,675 users, their personality types and a section of their last 50 posts on an online forum called personalitycafe.com. The online forum is a small community with around 153,000 members dedicated to discussing health, behavior, personality types and personality testing. As a result, the discussions are quite different to those on other social media platforms and we suspect the people using this online forum to be of a different demographic to those using other social media platforms. Consequently, the models utilising this dataset are likely not generalisable to data on other platforms like Facebook and Twitter. Like datasets discussed in Section 2.5.1, the Kaggle dataset is also very unbalanced – particularly in the E/I and N/S categories where Extroverts and Sensors account for less than a quarter of each respective class. The dataset is also relatively small, meaning machine learning and deep learning models could easily overfit the data if the number of trainable parameters aren't appropriately chosen. This dataset is the only known publicly available labelled dataset used for modelling the MBTI of social media users.

A number of academics have utilised this dataset to classify the MBTI of users through including natural language features on the information in their posts [15, 82, 6, 111]. Bharadwaj *et al.* [15] compared the results of a naive Bayes, SVM and MLP on this dataset with a set of 50 features. These features were created by performing a SVD on word counts, TF-IDF scores, LIWC scores and emotion based Scores (from EmoSenticNet and ConceptNet). The authors compared the accuracies with and without certain features and discovered the best results were achieved when all features were included. The best model was found to be the SVM classifier, with an average accuracy of 84.8%.

Keh and Cheng [82] also utilised the Kaggle dataset to predict the MBTI of users. These authors fine-tuned BERT's pre-trained language model for text classification and included four output layers, one for each trait. The authors compared the results for different epoch numbers and learning rates. They also considered the multi-classification problem by displaying how often the classifier got at least one, two, three and four personality dimensions correct. Their model achieved an average accuracy of 74.5% across the four personality dimensions.

Amirhosseini and Kazemian [6] compare an RNN and XGBoost model on the Kaggle dataset in predicting the MBTI of users. The authors begin by performing an analysis of the dependency structure of the four traits. They found that the four traits were fairly independent in this dataset, which has been assumed true by most other authors utilising the data. As a result the authors fit four independent XGBoost and RNN models. The authors find the XGBoost classifier significantly outperforms the RNN in the E/I, N/S and J/P traits, however the RNN performs better for the T/F trait. On average, the XGBoost classifier outperforms the RNN, with an average accuracy of 75.4%, compared to 67.8%.

Patil *et al.* [111] performed a a similar analysis on the Kaggle dataset. These authors performed a logistic regression, naive Bayes, SVM and random forest model on TF-IDF features from the data. In each case, they performed the multinomial model and compared the results to the multi-classification problem, rather than comparing each trait individually. They found achieved multi-classification accuracies of 64.38%, 45.63%, 64.88% and 61.42% for the logistic regression, naive Bayes, SVM and random forest models, respectively. Unlike previous authors, Patil *et al.* [111] reported confusion matrices for each of the models, where we can observe that some personality types receive no predictions. We also suspect this is the case with all previously discussed models on this dataset, as none of the authors appear to address the issues with class imbalance in the dataset. As a result, it is likely that any model not appropriately dealing with class imbalance will have accuracies which poorly represent model performance – particularly for the E/I and N/S traits.

Other authors have tried to model the personality type of users on other social media platforms [117, 59]. Plank and Hovy [117] model the MBTI of Twitter users through utilising a small dataset of 1,500 users and Gjurković and Šnajder [59] model the MBTI of Reddit users. To our knowledge, these are the only two papers which model the MBTI of users on data other than the Kaggle dataset.

Plank and Hovy [117] obtain their data by considering a corpus of 1.2M tweets from 1,500 users that self-identify with an MBTI over the duration of about one week. To find the self-identified users, the authors queried for the term 'Briggs' and returned all results over the considered timeframe. They then manually checked all files and removed all tweets mentioning more than one personality type (approx. 30%) – often these tweets refer to someone else's personality type or are bot posts. The authors then fit four independent Logistic Regression (LR) models to the data and consider the models with and without controlling for gender. They consider n-grams ($n = 1, 2, 3$), gender and Twitter features (followers, listed counts, etc) as their predictors in the model. They report the most predictive features under each model and discuss the direction of the relationship with respect to the dichotomy. The authors find that the N/S and J/P traits receive higher accuracies when controlling for gender, whereas the E/I and T/F traits do not. The average model accuracy when controlling for gender is 66.0% and the average model accuracy without controlling for gender is 66.6%.

Gjurković and Šnajder [59] obtain their data by firstly acquiring users who have any mention of a MBTI type in their flair field (sub-reddit bio). Of these users, they removed those who had ambiguous flairs, for instance people who mention more than one type. The authors then found some of the personality types to be underrepresented in the population and consequently achieved more samples of the underrepresented types by querying for mentions of these types in Reddit comments. They used variations of the query; "I am (an) <type>" and manually filtered ambiguous comments. What results was a dataset comprised of 22,934,193 comments from 13,631 unique users and 354,996 posts from 9,872 unique users, making it the largest labelled Myers-Briggs personality dataset to date. For their analysis, the authors create a subset of data which contained the comments of all users who contributed more than 1000 words to the dataset. The authors consider TF-IDF weighted n-grams, LIWC scores and MRC ratings as their features and model each of the personality traits independently. They consider the relevance of features by performing a t-test on every feature for each dichotomy. They then rank the importance of groups of features for each dichotomy by the proportion of significant features in each group. For instance, 41%, 29%, 62% and 40% of the LIWC features are relevant for the E/I, N/S, T/F and J/P traits, respectively. The authors compare the results of LR, SVM and MLP models on all the features and find a mixture of LR and MLP models work best based on the F1 score. They additionally report a F1 score and accuracy for the multi-classification problem, which received a F1 score of 41.7 and an accuracy of 69.2%. Unlike previously discussed models of the MBTI, Gjurković and Šnajder [59] also acknowledge the class imbalance problem and perform class weighting to remediate these issues – they also use a majority class classifier MCC as a comparison for all the results.

Table 2.3 provides an overview of the literature modelling the Myers-Briggs Personality type of users on social media services. Since authors utilise different metrics when assessing model performance and some model the traits altogether, it is difficult to compare the performance of the models. From Table 2.3, it appears the MLP and SVM models by Bharadwaj *et al.* [15] are achieving the best scores for the accuracy metric, however these authors do not address problems with class imbalance in the dataset and their models may resultantly have low model performance. Patil *et al.* [111] models the traits altogether and achieves a very good overall accuracy to the multi-classification problem. However, Gjurković and Šnajder [59] are the only authors who address class imbalance issues in their data and consequently perform weighted models in order to achieve metrics which accurately represent model performance. In most cases, it seems like the N/S trait achieves the highest values for considered metrics, with the E/I trait also achieving decent performance. Moreover, it appears that the J/P trait is the most difficult to model in most cases.

Authors	Data	Features	Model	Traits	Metric	Value
[15]	Kaggle	TF-IDF,LIWC,Emotion	MLP	E/I	Accuracy	85.3
[15]	Kaggle	TF-IDF,LIWC,Emotion	MLP	N/S	Accuracy	90.4
[15]	Kaggle	TF-IDF,LIWC,Emotion	SVM	T/F	Accuracy	87.0
[15]	Kaggle	TF-IDF,LIWC,Emotion	SVM	J/P	Accuracy	79.0
[82]	Kaggle	BERT	seq model	E/I	Accuracy	75.83
[82]	Kaggle	BERT	seq model	N/S	Accuracy	74.41
[82]	Kaggle	BERT	seq model	T/F	Accuracy	75.75
[82]	Kaggle	BERT	seq model	J/P	Accuracy	71.90
[6]	Kaggle	TF-IDF	XGBoost	E/I	Accuracy	78.17
[6]	Kaggle	TF-IDF	XGBoost	N/S	Accuracy	86.06
[6]	Kaggle	TF-IDF	RNN	T/F	Accuracy	77.8
[6]	Kaggle	TF-IDF	XGBoost	J/P	Accuracy	65.7
[111]	Kaggle	TF-IDF	XGBoost	All	Accuracy	64.88

[117]	Twitter	n-gram,Twitter features	LR	E/I	Accuracy	72.5
[117]	Twitter	n-gram,Twitter features	LR	N/S	Accuracy	79.5
[117]	Twitter	n-gram,Twitter features	LR	T/F	Accuracy	61.2
[117]	Twitter	n-gram,Twitter features	LR	J/P	Accuracy	58.2
[59]	Reddit	TF-IDF,LIWC,MRC	MLP	E/I	F1	82.8
[59]	Reddit	TF-IDF,LIWC,MRC	MLP	N/S	F1	79.2
[59]	Reddit	TF-IDF,LIWC,MRC	LR	T/F	F1	67.2
[59]	Reddit	TF-IDF,LIWC,MRC	LR	J/P	F1	74.8

Table 2.3: An overview of the literature to date which has modelled the Myers-Briggs personality type of social media users. We present the best models from each paper, as determined by the metric chosen by each of the authors. Most authors model each trait independently, however Patil *et al.* [111] model the traits altogether and do not assume independence. It is difficult to compare model performance as authors utilise different metrics during assessment. However, it appears that the best accuracies were achieved by Bharadwaj *et al.* [15], where they considered a large number of features in three different machine/deep learning models. Gjurković and Šnajder [59] were the only authors who acknowledged problems with class imbalance and still achieved good results using the F1 Metric for assessment.

OCEAN Traits	Myers-Briggs Traits			
	E/I	N/S	T/F	J/P
Openness	0.03	**-0.72**	0.02	0.30
Conscientiousness	0.08	0.15	-0.15	**-0.49**
Extroversion	**-0.74**	-0.10	0.19	0.15
Agreeableness	-0.03	-0.04	**0.44**	-0.06
Neuroticism	0.16	0.06	0.06	0.11

Table 2.4: Correlations between the OCEAN and Myers-Briggs personality traits. Values were calculated based on a study of 267 men and 201 women who undertook both personality tests. The results for each of the OCEAN traits were correlated with the Myers-Briggs traits, where the Extroverted, Intuitive, Thinking and Judging traits were represented negatively and their contraries were represented positively. Significant correlations are displayed in bold. These results were obtained from the paper produced by McCrae and Costa [95].

While the OCEAN and Myers-Briggs models were created through fairly different methods, the two models are still somewhat related [95]. McCrae and Costa [95] performed a study which considered how the traits in each of these models relate with one another. In particular, they performed a study with 267 men and 201 women to measure how peoples OCEAN type correlated with their Myers-Briggs type. They observed the correlations presented in Table 2.4, where we can see significant correlations with each of the Myers-Briggs traits. For instance, we observe that extroversion correlates with extroversion in both models, neuroticism correlates with openness and feeling correlates with agreeableness. However, it appears that Neuroticism is not well represented by the Myers-Briggs traits and consequently the authors suggested that Myers-Briggs measures four out of five major dimensions of personality.

2.5.3 The Detection and Influence of Bots

In this section, we will summarise some of the literature which has used models to understand how bots influence discussions on social media services.

Bots are by no means merely a technical phenomena, but they change the way internet users interact and form connections with each other. Only recently have researchers begun to address their potential to intervene in campaigns and to distort public opinion, communication and deliberation [83]. Much of the work in this area has focused on how to detect bot-like accounts on social media [109, 83, 38]. Orabi *et al.* [109] provided an systematic review of bot detection techniques and have shown bots are present in social networks, especially with regards to political campaigns/movements [109]. The authors discuss how bots have impacted significant movements and can alter the opinions of people online. However, the paper is simply just an overview of the techniques for detecting bots and does not present any new methods for detecting them or measuring their influence.

Pozzana and Ferrara [121] acknowledge that a lot of research has focused on bot detection, with little attention on the characterisation and measurement of bot activity and behavior. The paper studies the behavioral dynamics that bots exhibit over the course of an activity session, and highlight if and how these differ to human accounts. The authors measure the propensity of users to engage in social interactions or produce content. Their research indicates the presence of short-term behavioral trends in human accounts which are absent in bots.

Stella *et al.* [141] took more of a networks sciences approach on the problem and showed bots increase exposure to negative and inflammatory content in online social systems. These authors find that 19% of overall interactions are directed from bots to humans, mainly through retweets (74%) and mentions (25%). They consider the time series of tweet volume and bot activity over approximately 10 days during the Catalan referendum. They found that bots significantly increased their Twitter reply volume during the days of the referendum and deduced that bots preferred this form of targeted communication. The authors consider the sentiment directed between all pairwise combinations of bot accounts and human accounts. They observe a drop in sentiment in human-to-human and bot-to-human interactions on the day of the referendum. They make the important distinguishment between social and emotional interactions, where social interactions describe a retweet, reply or mention and emotional interactions is defined by the sentiment directed between accounts. The authors provide a network visualisation of both forms of interactions between accounts based on their alignment on the referendum.

Cresci *et al.* [38] used a more socially-focused approach. The authors then produce the first empirical evidence of Twitter spambots. They obtain datasets of verified human-operated accounts, social spambots, traditional spambots and fake follower account through a crowd-sourcing campaign. They first measure Twitter's current capabilities of detecting different types of bots. They find that Twitter suspends a majority of traditional spambots and fake follower accounts. However, they observe that Twitter fails in detecting and suspending social spambots in over 95% of cases. Moreover, the authors measure whether humans can succeed in detecting bots. They found that humans can detect social spambots in less than 24% of cases, traditional spambots in 91% of cases and human-operated accounts in 92% of cases. These results display the existence of striking differences between traditional and social spambots. More concerningly, they display that humans often can't distinguish between social spambots and human-operated accounts. Consequently, the presence of social spambots can go unnoticed without accurate bot-classification systems, making their presence on social media even more worrying.

Summary

In this chapter we have provided a background on the work to follow in the subsequent chapters. First we motivated research on social media platforms and how these provide an environment for which online digital footprints may be misused. We then discussed two popular personality models; the OCEAN model and the Myers-Briggs model. We presented the psychological framework of both models, discussed their application to various fields and considered their limitations. Next we presented the relevant NLP tools to be used in our methods and results. We then provide the binary mathematical models which we utilise in modelling personality types and the time-series models which we utilise in our analysis of the Russia/Ukraine conflict. We then perform a literature review of any work related to our research. In particular, we review the work which has focused on online personality profiling using both personality models. We also consider the work which has focused on the detection/influence of bots online. In the next chapter we will provide a methodological framework for modelling the personality types of users on Twitter. However, the framework will be generalised such that it can be applied to the labelling of users on any online service.

Chapter 3

Modeling Personality using Online Digital Footprints

Since the uprise of social media, personality profiling has become a very topical area in social media research and natural language processing [143]. It has been utilised by various companies and governments for applications such as targeted advertising, political campaigns and vaccine campaigns. A high-profile example of its impactful implementation is the example of Cambridge Analytica, a British political consulting firm which are believed to have used personality profiling to impact the results of the 2016 US election as well as the withdrawal of the United Kingdom from the European Union [162, 68]. It is suggested that useable information can be obtained about someone by inferring their personality type, including personal attributes such as values, emotional stability or how they behave in relationships [162]. Since this data can be misused, it is of interest for individuals to understand the extent of information that is attainable from their online digital footprint. This is also of key concern for governments, who seek to maintain democracies and the ethical use of data, both of which can be attacked through the use of personality profiling by companies like Cambridge Analytica [162].

We aim to explore the extent to which someone's online digital footprint can be utilised to model their personality type. Firstly, we collate a dataset of approximately 44,000 accounts self-labelled with their Myers-Briggs personality types on Twitter – this is the largest labelled Twitter dataset of personality types (to our knowledge). We discuss various biases which may arise through the processes we use in collecting this dataset and consider the independence of the personality traits. Moreover, we produce various linguistic, sentiment and network based features for each user, where we aim to use them in our personality models. We then develop a statistical framework to train and identify the most appropriate machine learning model for predicting the Myers-Briggs personality type of Twitter users. We consider a broad suite of different weighting/sampling techniques which help model unbalanced data, something we observe in our dataset. We test these weighting/sampling techniques on our machine learning models, consisting of: logistic regression, naive Bayes, support vector machines and random forests. While we develop this statistical framework to model personality types, the framework can also be applied more generally to any group of models and any labelled characteristic about an account. This may include an accounts political opinion, their psychological properties or even someone's propensity to adopt an opinion/viewpoint. Finally, we develop a method for analysing the importance of individual features and groups of features in our models. We use this method to explore how personality types use language differently, and discover the most informative features in our models. Our analysis in this chapter aims to provide individuals with an understanding of how their online digital footprints can be used to understand personal qualities about them, highlighting the importance of data privacy. In essence, we explore the performance of personality profiling models and discuss the extent to which they could be misused.

3.1 Data

In this Section, we explain how we collected the labelled personality dataset used in our analysis. We then describe the preprocessing steps used to clean the data and prepare it for the machine learning models. Finally, we perform an exploratory data analysis (EDA) on the dataset to describe the important features and any problems which may arise with modelling the data.

3.1.1 Data Collection

Data collection is an important part of this work as it enables us to obtain a labelled dataset for use in supervised machine learning models. Early exploration of Twitter led us to find that a number of online accounts would self-report labelled attributes about themselves. These labelled attributes included information such as people's star-signs, political opinions and personality types. We focus on the self-reporting of personality types, since personality profiling is a current focus of social media research and can often be linked to other attributes about a person such as their political opinions [56]. We collect the data for this analysis by utilising the academic Twitter API (V2). Note that while we analyse users from Twitter and chose personality types as the attribute about a person we wish to model, our subsequent methods are generalisable to any social media platform and any type of labelling for an account. Our exploration of Twitter identified that accounts would almost always report categorical personality traits, rather than raw scores from a personality test. For instance, accounts would often report one of the sixteen four-letter acronyms from the MBTI or one of the nine Enneagram[1] types, but accounts rarely report their numerical scores for each trait from the MBTI or OCEAN models.

[1] A system of personality which describes people in terms of nine types, each with their own motivations, fears, and internal dynamics.

Unlike the MBTI, OCEAN personality traits are not represented with dichotomies and consequently respondents don't obtain acronyms representing their personality type. These four-letter acronyms give people a short categorisation of their personality that is easily self-reported on social media in the form of a regular expression. As a result, people are much more likely to self-report their categorical MBTI type rather than their OCEAN personality type. The popularity and familiarity of the MBTI among the general population also means people will self-report MBTI results more often on Twitter. The four letter MBTI acronyms are also unique to the Myers-Briggs questionnaire, meaning they can be easily queried using the Twitter API. This also means these personality types won't be confused with any other acronym or word, reducing the likelihood we incorrectly classify any users. When we initially explored Twitter, we found that some users self-reported their personality type in their biography and other users would self-report their personality types in their tweets. As a result, we formulated two methods for querying and labelling the Myers-Briggs personality type of accounts. Firstly, let Ω define the set of 16 acronyms for Myers-Briggs personality types, as seen in Figure 2.2. We describe the two methods below:

- **M1** Search query: $\{\mathtt{x} : \mathtt{x} \in \Omega\}$. We provided this query to the Tweepy's `search_users` endpoint to obtain the set of users who currently self-report their personality type in their username or biography. Due to the rate limits associated with this endpoint we were limited to obtaining no more than 1000 users for each unique search query.

- **M2** Search query: $\{(\text{I am } \mathtt{x}) \vee (\text{I am a } \mathtt{x}) \vee (\text{I am an } \mathtt{x}) : \mathtt{x} \in \Omega\}$. We then provided this query to the Twitter API's `full_archive_search` endpoint to obtain the set of users who have self-reported their personality type in a Tweet since Twitter's creation (March 26, 2006). Note that we only searched for self-reports in Tweets and excluded Retweets, Quotes and Replies in our query due to a much higher potential of incorrectly labelling an account. Furthermore, we were bound by rate limits of 300 requests per 15-minute window, however there were no hard bounds on the number of tweets or users we could obtain. As a result, we ran this query for each personality type until the search was exhausted.

Note that in both cases, the queries were not case-sensitive.

Using both methods **M1** and **M2**, we were able to obtain a dataset comprising of 68,958 users with a labelled Myers-Briggs type. To determine the percentage of accounts which were mislabelled for each method, we randomly sampled 1000 accounts from each dataset and manually checked them. We display the number of accounts obtained and the mislabelling rate for both methods in Table 3.1.

The mislabelling rate was higher for **M2** compared to **M1**. Often accounts were mislabelled from **M1** because of two reasons: either bot-like accounts had a automated username which just happened to reference one of the four-letter acronyms or accounts were referencing personality types other than their own. One example of the former was an

	M1	**M2**
Accounts	15,986	52,972
Mislabelling Rate	1.9%	3.4%

Table 3.1: Summary of the accounts obtained for each search query method. We include the total number of accounts and the mislabelling rate. The mislabelling rate is determined by manually checking the accounts.

account with the username 'ISTJ093' and one example of the latter was an account with the following text in their biography 'I'm here to endorse Female ISTP rights'. Whereas, accounts were often mislabelled from **M2** because either: the account was referencing personality types other than their own or the account was referencing more than one personality type. One example of the former was an account which tweeted 'I don't think I am an ENTP...' and one example of the latter was an account which tweeted 'i used to be infp, i am now intp...'. As a result of misclassifications like these, we introduced a number of preprocessing steps and an inclusion-exclusion criteria which minimised the number of accounts which were potentially mislabelled during data collection. These will be discussed in Section 3.1.2.

As part of our data collection procedure, we were required to obtain information about the accounts which could be utilised as features in our machine learning models. We used the philosophy that 'any extra information could be useful information' and consequently collected a substantial amount of data for every account. We began by using the Twitter API's `full_archive_search` endpoint to collect each accounts most recent 100 tweets or quotes. Note that when obtaining quotes we only returned the text from quote itself and not any text from what's actually being quoted – this is so we only analyse text written by the labelled account. We included tweets and retweets in the analysis because they are both directed to the accounts followers. A tweet gives us information about how an account shares experiences and opinions, whereas a quote gives us information about how an account responds to other content. We excluded retweets from this set of data because retweets are not written by the labelled account and thus contain little to no information about the accounts use of linguistic features. Moreover, we excluded replies because they are generally much shorter and directed at an individual rather than an accounts followers.

Next we utilised Tweepy's `lookup_users` endpoint to obtain the account characteristics for each user. This included the biography attached to every account as well as a set of Social Metadata (SM) features. The user's biography and the 100 tweets/quotes were used to generate a set of linguistic features – this will be further described in Section 3.1.2. Whereas, the SM features will be directly used as numeric features in the models and these are described in Table 3.2.

SM Feature	Description
Followers Count	The number of followers this account currently has.
Friends Count	The number of users this account is following (AKA their "followings").
Listed Count	The number of public lists that this user is a member of.
Favourites Count	The number of Tweets this user has liked in the account's lifetime.
Geo Enabled	When true, indicates the the user has attached geographical data.
Verified	When true, indicates that the user has a verified account.
Statuses Count	The number of Tweets (including retweets) issued by the user.
Default Profile	When true, indicates that the user has not altered the theme or background of their user profile.
Default Profile Image	When true, indicates that the user has not uploaded their own profile image and a default image is used instead.
Profile use Background Image	When true, indicates a user has added their own background image to their account.
Has Extended Profile	When true, indicates a user has attached an extended profile link to their profile.

Table 3.2: Summary of Social Metadata (SM) features which we obtain using the Twitter API and utilise in our machine learning models.

3.1.2 Preprocessing

After collecting the accounts, it was important to perform a number of preprocessing steps on the data to prepare it for modelling and analysis. During these preprocessing steps we utilised a number of NLP toolkits in Python, a majority of which are described in Section 2.3. We perform preprocessing with the purpose of maximising the performance of our models on the data. These preprocessing steps can be separated into two parts: data cleaning and feature extraction.

It was crucial to correctly format and clean the data in preparation for linguistic feature extraction. The linguistic features we considered in our models can be separated into four categories: LIWC features, BERT features, VADER features and Botometer features. We call these linguistic features because they are numerical representations based on a user's text usage, whereas the Social Metadata (SM) features in Table 3.2 are based on the characteristics of a user's account. We distinguish between different types of linguistic features because the data cleaning steps in preparation for their extraction is dependent on the feature type. For instance, the linguistic data cleaning steps we are about to cover are in preparation for LIWC feature extraction. Whereas, the BERT, VADER and Botometer features required no linguistic data cleaning and can be obtained through utilising raw text directly from the Twitter API.

The first step of our data cleaning processes was removing any instances of duplicate users. We did this by only keeping the users latest mention of a personality type. Note that tweets were assigned the UTC time they were tweeted and biographies were assigned to be the current UTC time. As a result, we always considered biographies to be the most accurate representation of a users current personality type, rather than tweets. This is because there is a lower misclassification rate for accounts queried based on their biography (see Table 3.1).

We then performed the same linguistic data cleaning techniques on the biography and tweets by combining them into a single string for every account – we refer to this as the combined text. The remaining data cleaning steps were performed to prepare the data for LIWC feature extraction. These steps are outlined below and were all performed on the combined text:

1. Normalised the text using the Normalization Form Compatibility Decomposition (NFKC) algorithm[2].

2. Calculated each accounts predominate language through utilising Spacy's Language Detector.

3. Removed all non-English language from the text using the Python bindings for the Compact Language Detect 2 (PyCLD2) library.

4. Calculated the language dependent Botometer scores for the accounts with English as their predominate language.

5. Converted the text to lowercase.

6. Removed URLs, email addresses, punctuation and numbers.

7. Tokenized the text using the Tweet Tokenizer from the Natural Language Toolkit (NLTK) [16].

[2] where characters are decomposed by compatibility, then recomposed by canonical equivalence.

8. Removed any empty tokens and removed any mention of one of the 16 MBTI acronyms.

The choice and order of these techniques were decided after an exploratory study. Before performing any data cleaning, we observed the formatting of words in the LIWC dictionary and discovered nuances in the composition of different words. For instance, the text was always normalized, in lowercase and in English. There were no instances of URLs, email addresses or numbers, and punctuation was always removed, except for when used in emoticons. As a result, a number of the data cleaning techniques we used were necessary to ensure words from the LIWC dictionary were detected. However, we often had the choice of what NLP toolkit we wanted to use to perform the data cleaning. Spacy's language detector was used to calculate the predominate language for each user, as it performed well on a random sample of tweets [3]. The PyCLD2 library was chosen as it was the only open source Python library (to our knowledge) which could detect the use of multiple languages in a string and return the positions for which English language is used. Botometer was used to calculate bot scores because it is the most widely used API which calculates bot scores on Twitter accounts. We utilised the language dependent scores from Botometer because previous literature found these to give more accurate results for English speaking accounts [163]. The Tweet Tokenizer by NLTK was used as it is the most commonly used tokenizer for performing any tokenizing tasks on text from Twitter. While removing punctuation was an imperative choice to ensure words such as 'what's' were recognised by LIWC, it also meant that emoticons were removed. We justify this choice because there were significantly more tokens with punctuation than emoticons.

Next, we formulated an inclusion-exclusion criteria to determine whether a personality could be profiled from a Twitter account. We formulated four basic conditions which aimed to ensure an account's data would be informative of their personality without removing too many accounts from the dataset. Firstly, we wanted to ensure the users had enough textual data so that linguistic features were representative of an account's language use. We note that Keh and Cheng [82], Amirhosseini and Kazemian [6], Patil *et al.* [111] achieved acceptable results using each accounts most recent 50 posts to a personality forum, as we discussed in Section 2.5.2. However, the posts on the personality forum were genuinely longer than the tweets we observed in our dataset. As a result, we decided a minimum requirement of 100 tweets per account was suitable to ensure there was enough linguistic content, without removing too many accounts from the dataset.

The LIWC, BERT, VADER and Botometer features are all determined based on only English language, so it was an obvious choice to only consider English speaking accounts in our inclusion-exclusion criteria.

Next we considered the number of bot accounts which might be present in our dataset. As previously mentioned, one reason for misclassifying personality types was due to bot-like accounts using automated usernames which happened to reference one of the four-letter MBTI acronyms. Consequently, we deduced it was appropriate to include a condition in our inclusion-exclusion criteria which minimised the presence of bots in our dataset. We did this by defining a manual threshold which utilises the Botometer CAP score to classify bots in a binary fashion. Note that we use the Botometer CAP score because we are interested in the overall bot likelihood and not the sub-category bot likelihoods. There seems to be no consistency in the literature surrounding what threshold is best for binary classification. Rather, authors define their threshold based on a false positive rate which makes sense for their problem. For instance, Wojcik *et al.* [160] use a threshold of 0.43 for their political analysis of the twittersphere, whereas Keller and Klinger [83] use a larger threshold of 0.76 for their analysis of social bots in election campaigns. In our analysis, we wanted to avoid large numbers of false positive bot classifications due to the fairly limited number of accounts in our dataset – so we opted for a threshold of 0.8. This larger threshold meant we would correctly classify accounts which were almost certainly bots but also meant it was likely that we would falsely classify some bots as humans. Similarly, Giorgi *et al.* [58] indicated that bots also express a personality which exhibits very human-like attributes. So even if we falsely classify a small number of bots, they are still likely to exhibit a personality profile.

Finally, we considered the number of MBTI types which an account mentioned in its tweets. We previously noted that one reason for the misclassification of accounts was users who mentioned more than one type in their tweets. This meant that these users often appeared more than once in our dataset before the removal of duplicate users. While the removal of duplicate users meant these accounts could only be assigned one type, it also meant that accounts which referenced two different types could potentially be misclassified. We therefore removed all accounts which referenced more than one type in their tweets. This meant that we removed ambiguity in the labelling of accounts in our data, further reducing the mislabelling rate of accounts. In Table 3.3 we present the four conditions which constitute the inclusion-exclusion criteria for Twitter accounts to be personality profiled. After removing the accounts which did not satisfy these four conditions, there were a total of 43,977 accounts remaining. These accounts then formed the users in our labelled dataset which we used to train and test our models.

Quality	Condition
Tweets/Quotes	> 100
Predominate Language	= English
Botometer CAP	< 0.8
MBTI Types Referenced	$= 1$

Table 3.3: Inclusion-exclusion criteria for a Twitter account to be personality profiled using our models.

The final part of our preprocessing steps on the data was to extract the LIWC, BERT and VADER features from the text. We noted above that the data cleaning techniques were performed for LIWC feature extraction, whereas the BERT and VADER features can be extracted directly from Twitter's raw text output. Consequently, we used the processed tokens from the Tweet Tokenizer to calculate the LIWC features. Suppose we wish to calculate the LIWC features for some user, u, in our dataset with n_i tokens in their i^{th} most recent tweet. Furthermore, let $t_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,n_i}\}$ denote the set of tokens in the i^{th} most recent tweet and $N = \sum_{i=1}^{100} n_i$ denote the total number of tokens for the user. We calculate the LIWC features by micro-averaging the tokens which are present in each category. Hence, the score $s\left(c\right)$ for some category, c would be:

$$s\left(c\right) = \frac{100}{N} \sum_{i=1}^{100} \sum_{j=1}^{n_i} \mathbb{1}_{\{w_{i,j} \in c\}}.$$

Recall that these words are mapped to LIWC categories via a one-to-many mapping, as we discussed in Section 2.3.1. Hence, these scores will not necessarily sum to 100, but can rather be interpreted as the proportion of a user's tokens which are contained in each LIWC category. In this case, we micro-average the results per token because we value the frequency of someone's words to be more important than the context of their tweets. We justify this with a result from Pennebaker and Francis [114], who discovered that someone's use of highly frequent words (such as pronouns like 'I', 'you' and 'we') says more about their personality than the context of what they actually write. While this result was found from controlled psychological experiments, we will further discuss its importance to our models in the results for this chapter.

Next we calculate the BERT features on the raw Twitter output for each user. We do this by using the base BERTweet language model, discussed in Section 2.3.3. Recall that we use BERTweet's in-built tokenizer, so the tokens are likely different to those used when calculating LIWC features. For brevity, we use the same notation as above to define tokens and tweets. We calculate the BERT features by firstly obtaining a single embedding vector for each tweet, and secondly obtaining a single embedding vector for each user. We calculate the embedding vector for each tweet by averaging the embeddings for each token:

$$E_{t_i} = \frac{1}{n_i} \left(E_{i,1} + E_{i,2} + \cdots + E_{i,n_i}\right),$$

where $E_{i,j}$ denotes the 768-dimensional embedding vector for the j^{th} token in the user's i^{th} most recent tweet. We then calculate the embedding vector for each user by averaging the embedding vectors for each tweet:

$$E_u = \frac{1}{100} \left(E_{t_1} + E_{t_2} + \cdots + E_{t_{100}}\right).$$

The 768 elements of E_u then become features in our models and aim to represent the average context of the user's tweets. In this case, we chose to macro-average the tokens present in each tweet, rather than micro-averaging them like we did with the LIWC features. Recall BERT measures the context of text, compared to LIWC which focuses on the frequency of categories in the text. We would expect the context of words inside a tweet to be fairly similar, but the context of tweets to often differ. By macro-averaging the embeddings, we obtain a measure of the average context per tweet. We do this because we value the context of each tweet equally; a tweet of few words can describe a context in as much detail as a tweet with many words [134]. Another reason for doing this is that we already micro-average the proportion of words in each category with LIWC, so macro-averaging the results from BERT will potentially introduce less redundant information to our models.

Finally, we calculate the VADER features on the raw Twitter output for each user. Recall that VADER provides four scores for every input of text. These outputs are the proportion of positive, negative and neutral words as well as the overall sentiment of the text, as discussed in Section 2.3.2. The positive, negative and neutral proportions always sum to one, so we exclude the neutral proportions, as they would be redundant in our models. We calculate two versions of the remaining three VADER features: one set for a user's tweets and another set for a user's biography. We do this for two reasons: (i) the context of both strings of text; a user's biography most often contains information about themselves, whereas a user's tweets often contain information about their environment. Hence, the sentiment of a user's biography will often describe how they feel about themself and the sentiment of a user's tweets will describe their feelings towards their environment. (ii) it is not so costly or computationally expensive for us to do so. Separating these features only adds three more variables to the model; this is considerably less than other groups of features like LIWC or BERT which would add 74 or 768 features to the model, respectively. VADER features can also be calculated within a few seconds for each user, making them computationally efficient compared to other features like BERT. As a result, we have collected a total of 866 features comprised of:

- 11 SM features,
- 7 features from Botometer,
- 74 features from LIWC,
- 768 features from BERT, and
- 6 features from VADER.

In Table 3.4, we present the linguistic features which are used in our models. These features, combined with the SM features in Table 3.2 form the total feature set for our models.

Category	Features
Botometer	cap_english, english_astroturf, english_fake_follower, english_financial, english_other, english_self_declared, english_spammer
LIWC	function, pronoun, ppron, i, we, you, shehe, they, ipron, article, prep, auxverb, adverb, conj, negate, verb, adj, compare, interrog, number, quant, affect, posemo, negemo, anx, anger, sad, social, family, friend, female, male, cogproc, insight, cause, discrep, tentat, certain, differ, percept, see, hear, feel, bio, body, health, sexual, ingest, drives, affiliation, achiev, power, reward, risk, focuspast, focuspresent, focusfuture, relativ, motion, space, time, work, leisure, home, money, relig, death, informal, swear, netspeak, assent, nonflu, filler, total_word_count
BERT	$\{e_i\ ;\ i = 1, \ldots, 768\}$
VADER	tweets_sentiment, bio_sentiment, tweets_pos_words, bio_pos_words, tweets_neg_words, bio_neg_words

Table 3.4: Linguistic features in our model, separated by the feature type. Botometer features are language-dependent features and include the CAP score and sub-category scores. LIWC features are included based on how they appear in the LIWC dictionary – a full description of what each of these features represents, along with examples is provided in the paper by Pennebaker *et al.* [113]. BERT features include the 768-dimensional embedding vectors for each user and the VADER features include those calculated on both the biography and the tweets.

3.1.3 Exploratory Data Analysis

The Balance and Independence of Personality Types in the Dataset

We performed an exploratory data analysis (EDA) on the dataset to determine important information about our dataset prior to any modelling. As a first step, we wanted to consider how balanced our dataset was. We did this via two methods: firstly, we considered the balance of the 16 Myers-Briggs Personality Types; secondly, we considered the balance of each of the four dichotomies. When considering the balance of the 16 types, we wanted to highlight any biases which potentially arose in our data collection and pre-processing steps. We acknowledge and discuss two forms of potential bias in our dataset: firstly, bias which may arise from only considering MBTI types on Twitter, and secondly, bias which may arise from only selecting accounts which satisfy our inclusion-exclusion criteria as well as self-report their MBTI types on Twitter. We demonstrate these biases by producing two sets of bar plots in Figures 3.1 and 3.2. The former figure shows the proportions of MBTI types and the latter figure shows the proportions of the dichotomies – in each case, we display the proportions in our dataset, on Twitter and in the general population. We obtained the results from a study by Schaubhut *et al.* [136] to report the proportions of types present on Twitter. The authors of this study performed a survey on 1,784 participants from the US which involved an MBTI questionnaire and asked the users if they have a Twitter account. While this survey only considered US participants, it still provided us with a baseline to discover whether our sampling methods biased corresponding proportions in our dataset. Furthermore, we obtain the proportion of personality types in the general population from Robinson [129]. These authors utilised results from the official Myers-Briggs website (`https://www.myersbriggs.org/`), where data is sourced from users opting to sit their online questionnaire. While we refer to these results as the 'general population' in Figure 3.1, we also acknowledge that there may be biases in types which are more likely to sit the online questionnaire.

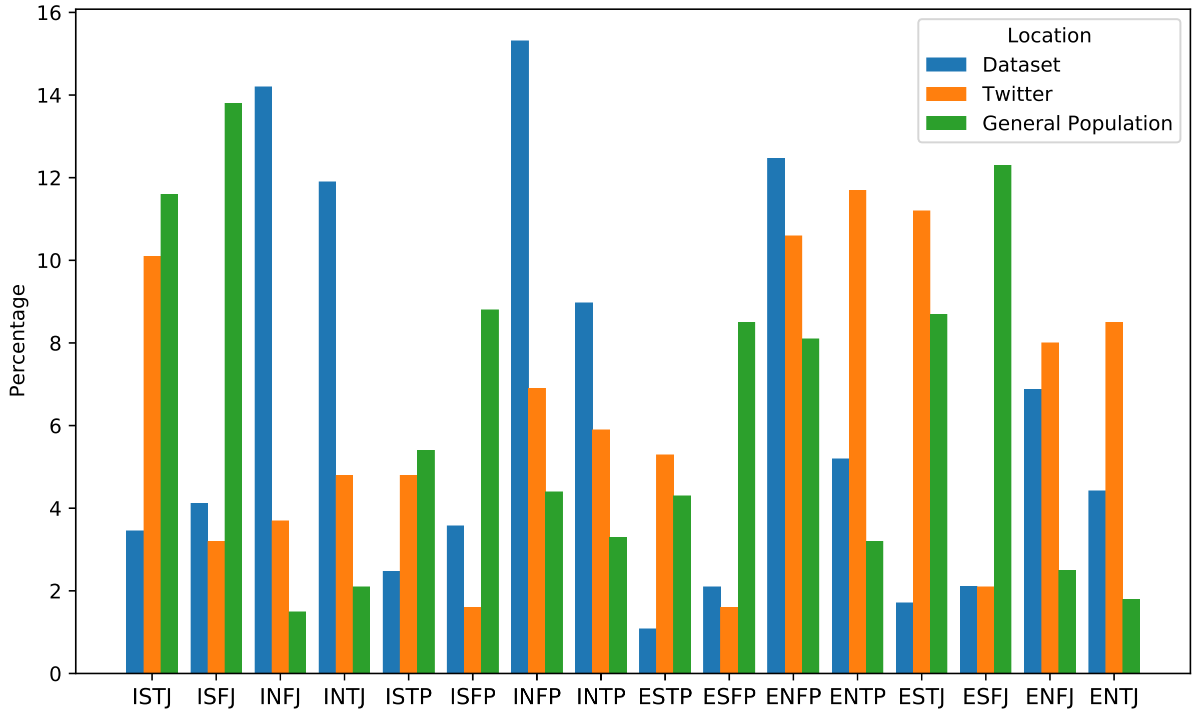


Figure 3.1: Proportions of the 16 MBTI types in our dataset, on Twitter and in the general population. We obtain the personality types on Twitter from a study by Schaubhut *et al.* [136] on 1,784 Americans. This study involved these participants sitting a MBTI questionnaire and answering whether they have an active Twitter account. Moreover, we determine the proportion of personality types in the general population from results by Robinson [129] which are obtained from the official Myers-Briggs website.

The first thing to observe in Figure 3.1 is the substantial class imbalances between

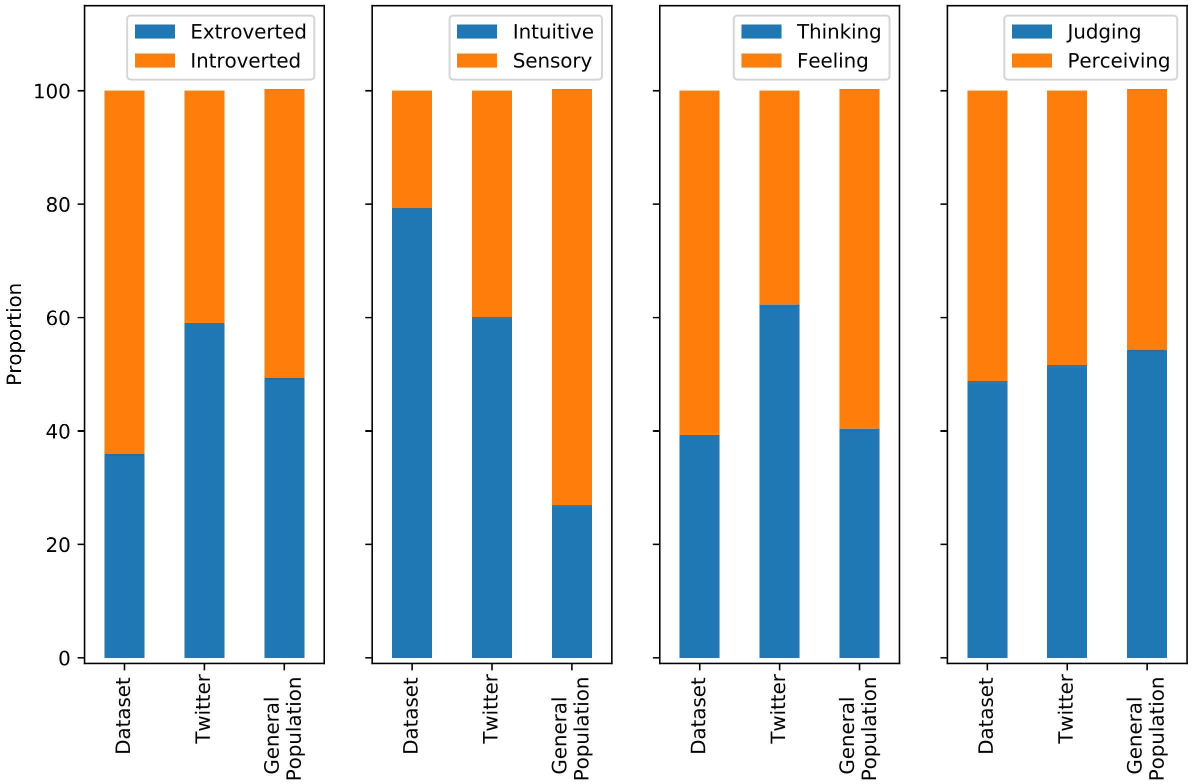


Figure 3.2: Proportion of accounts displaying each dichotomous trait in our dataset, on Twitter and in the general population.

personality types from the same location. This indicates that some personality types are more common than others, and humans generally have a tendency to exhibit certain personality traits over others. Moreover, we observe substantial differences between the proportion of the same types in different locations. For example, there are considerable differences in the proportions of ENTJ's and INTP's in our dataset, the Twitter dataset and the general population. This indicates that the two forms of bias are present in our dataset: (i) from only considering MBTI types on Twitter, and (ii) from only selecting accounts which satisfy our inclusion-exclusion criteria as well as self-report their MBTI types on Twitter. The biases from (i) are well represented by the differences in proportions between the Twitter dataset and the general population. Whereas the biases from (ii) are well represented by the differences in proportions between our dataset and the Twitter dataset.

While we see at least one form of bias is present for each type, it also appears that some types are not affected by both forms of bias. For instance, the proportions of some types in our dataset and on Twitter are barely separable (i.e. ESFP or ESFJ). The same is true for the proportions of some types on Twitter and in the general population (i.e. ISTP or ESTP). As a result, both forms of bias do not affect the proportions of every personality type in each location. Although, it is important to note that the proportion of each type in our dataset is always substantially different to the proportion of each type in the general population. This indicates that there is always at least one form of bias affecting each personality type in our dataset. However, bias was very difficult to avoid in this case. The bias introduced by considering accounts on Twitter and from our inclusion-exclusion criteria is not so worrying, because our models would only be directly applied to accounts on Twitter which satisfy our inclusion-exclusion criteria. Nonetheless, the same techniques and methodologies could be used to model personality types on other phenomena like Facebook, Reddit or in the general population, but these would require a different dataset specific to the platform of users being modelled. The bias introduced from our sampling of accounts which self-report their MBTI type is potentially more significant. This is because we will design a model with the intention to use it on Twitter accounts which don't self-report their MBTI type. We should note that our dataset may be biased from the Twitter dataset because of the inclusion-exclusion criteria as well as the requirement of users self-reporting MBTI types. Hence, it is likely that the inclusion-exclusion criteria also introduces some of this bias in our dataset. In any case, being able to obtain this dataset through appropriately querying MBTI types was unavoidable without huge amounts of funding to pay thousands of Twitter users to sit a long and cumbersome Myers-Briggs Questionnaire.

There is a very noticeable imbalance in the Intuitive/Sensory dichotomy across all datasets in Figure 3.2. Adding to this, there are observable imbalances in the Extrovert/Introvert dichotomy and the Thinking/Feeling dichotomy. Whereas, the Judging/Perceiving dichotomy is much more balanced across each dataset than the other dichotomies. Moreover, the imbalances in our dataset are mostly consistent with imbalances in the dataset obtained from `www.personalitycafe.com`. The higher proportion of introverts in our dataset is consistent with past research which found that introverts tend to use social media as a primary form of communication, whereas extroverts tend to prefer communicating in-person [72]. The larger proportion of intuitives in our dataset is consistent with Schaubhut *et al.* [136] which discovered that more individuals with a preference for Intuition (13%) reported being active users of Twitter than individuals with a preference for Sensing (8%). The same authors also found that intuitives spend more time browsing and and interacting with others on Twitter. The imbalance in the Thinking/Feeling dichotomy in our dataset is opposite to what we observe in the Twitter dataset. However, Schaubhut *et al.* [136] found that people displaying the Feeling trait are more likely to spend their personal time browsing, interacting and sharing information on Facebook. Provided the same is true for Twitter users, our inclusion-exclusion condition requiring users to be active on Twitter (i.e. tweet/quote at least 100 times) may bias our dataset leading to more users exerting the Feelings trait. Another interesting observation is the vast differences in the Intuitive/Sensory proportions for our dataset, the Twitter data and the data for the general population. The proportions in our dataset are strongly in favour of Intuitives, compared to the general population which is strongly in favour of Sensors, and the Twitter dataset lies inbetween the two. These differences may arise due to two biases: (i) MBTI types on Twitter, and (ii) the inclusion-exclusion criteria as well as requiring users to self-report their MBTI types on Twitter.

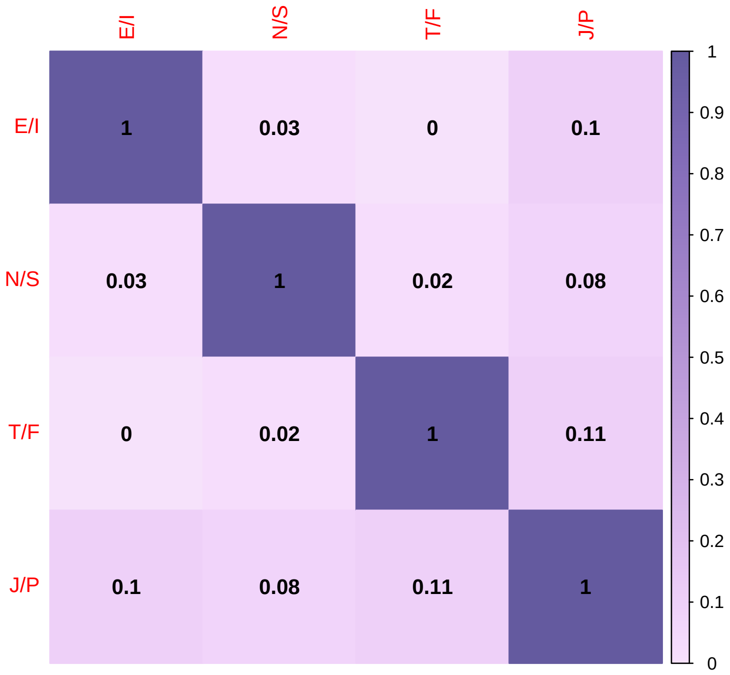


Figure 3.3: Pairwise results of the Cramér's V Statistic between each of the Myers-Briggs dichotomies for our dataset. Each panel displays and is coloured by the value of the Cramér's V Statistic. Note we use the bias corrected version of the statistic which is discussed in Section 2.4.2.

When modelling personality traits, some authors opt to classify the unique personality type of accounts and don't assume independence between the dichotomies [12, 111]. Whereas, most authors choose to model the dichotomies independently [1, 142, 15, 82, 6]. Often these authors show no justification for modelling the traits independently, which is a crucial assumption of the model. We determined the dependency structure of the four MBTI dichotomies in our dataset. We did this by utilising the bias-corrected version of the Cramér's V Statistic, which measures the association between two categorical variables, as discussed in Section 2.4.2. We calculated the Cramér's V Statistic on pairwise values from the MBTI dichotomies and display them in Figure 3.3.

The Cramér's V statistic is small in each combination of different dichotomies. As a result, we deduce that the four Myers-Briggs dichotomies are independent in our dataset. We emphasise that these results only hold for our dataset and cannot be extended to the general population or datasets from other social media services. However, the purpose of creating these models is to apply them to Twitter users who do not self-report their MBTI type. By doing this, we assume the dichotomies to be independent across the entire Twittersphere. Similarly, Amirhosseini and Kazemian [6] show the independence of the dichotomies on other social media services and Schaubhut *et al.* [136] show the proportion of personality types barely differs between social media sites. From these two findings, we expect the dichotomies to be independent across the entire Twittersphere.

Feature Independence in the Dataset

Another important aspect of our EDA was considering the features in our models. As part of this, we wanted to observe the relationships within and between the features. It is important to consider the relationships between these features to ensure there is no multicollinearity between features in our model – particularly for models like logistic regression. Since we have a total of 866 different features, it is not practical to display the relationships between every feature. Rather, we considered the relationship within and between the groups of features displayed in Tables 3.2 and 3.4. First, we display the correlations between all individual features, excluding the BERT features. We exclude the BERT features because; firstly, they do not provide any interpretable information about a user. BERT features are a vector representation of the semantic meaning of words – these numerical features are not designed to be mapped to a specific context. Secondly, there is simply just too many BERT features to meaningfully display. The resulting correlations are displayed in Figure 3.4, where we colour the axis labels based on the feature group they belong to.

While we observe some features to be correlated in Figure 3.4, most display weak correlations with other features. This indicates that we are unlikely to encounter problems with multicollinearity in our models. Moreover, we observe a majority of the stronger correlations to occur within the same feature group, rather than between different feature groups. For instance, we see a number of stronger positive correlations between each of the Botometer scores, but these features have significantly weaker correlations with features in other groups. It is likely this is due to 'bot-like' accounts displaying similar characteristics, regardless of their bot type. We observe similar dependencies for the LIWC features, where a number of them have relatively strong positive correlations with other features within the same group. This occurs because of the one-to-many mapping for words when calculating the LIWC scores for a user. Since most words appear in more than one category, we would expect categories containing the same words to be positively

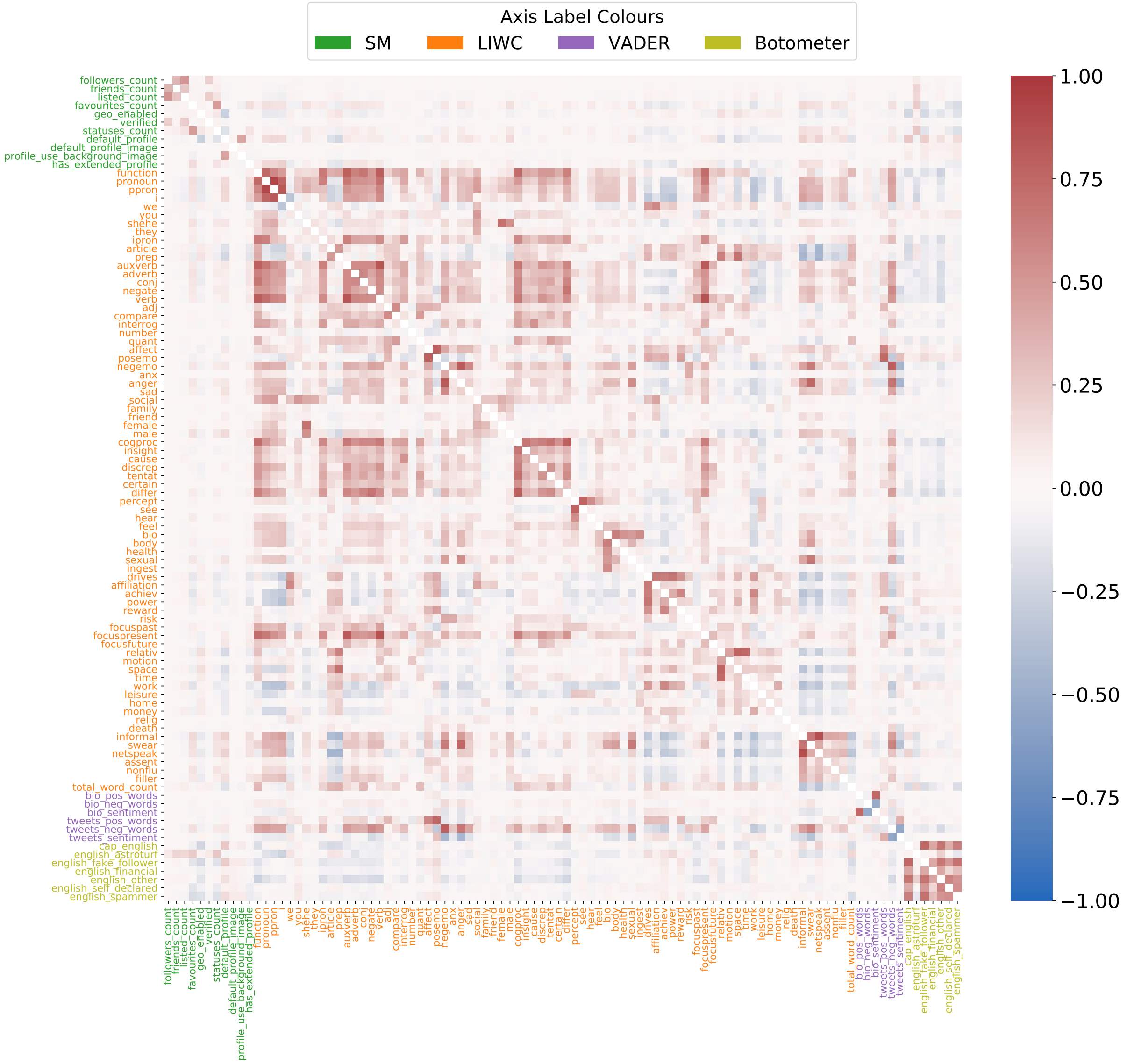


Figure 3.4: Correlations between the SM, LIWC, VADER and Botometer features. The axis labels are coloured based on the feature group they belong to, as described in the legend at the top of the figure. Each panel is coloured based on the pairwise Pearson Correlation between the two features, as described by the colour bar on the right hand side of the figure. Note that we exclude BERT features because they do not provide any labelled information about a user and there is simply just too many of them to meaningfully display.

correlated.

Next we considered the strength of correlations between different groups of features. We represented these correlations on a group level so we could visualise the dependencies within each group. We also included the BERT features in this analysis. Firstly, define $\alpha_{i,n}$ to denote the n^{th} feature in feature group i and let N_i denote the number of features in group i. We calculate the correlation between feature groups, i and j as following:

$$\hat{\text{cor}}\,(i,j) = \begin{cases} \frac{1}{N_i(N_i-1)} \sum\limits_{k=1}^{N_i} \sum\limits_{\ell=1,\ell\neq k}^{N_i} \left|\text{cor}\,(\alpha_{i,k},\alpha_{i,\ell})\right|, & \text{if } i == j, \\ \frac{1}{N_i N_j} \sum\limits_{k=1}^{N_i} \sum\limits_{\ell=1}^{N_j} \left|\text{cor}\,(\alpha_{i,k},\alpha_{j,\ell})\right|, & \text{if } i \neq j, \end{cases} \tag{3.1}$$

where cor () denotes the Pearson correlation between two features. Hence, we are affectively just averaging the correlations within and between different feature groups. Note that we do not include correlations between the same individual features because that will always be equal to one. Moreover, we take the absolute value because we want a measure of correlation strength and this avoids negative correlations cancelling positive correlations in the sum. The results between each of the different feature groups are displayed in Figure 3.5.

The average absolute correlation values in Figure 3.5 show a fairly weak correlation strength (≤ 0.11) between different groups of features. This indicates that the features are not multicollinear on a group level. It also highlights that these features provide fairly different information about a user. Prior to this, it was of concern that the LIWC and BERT features may be multicollinear because both sets of features describe the context of a users language use. However, Figure 3.5 has enabled us to verify that this is not the case and both of these groups of features provide different information to our models. Moreover, the average absolute correlations along the diagonal in Figure 3.5 show there to be stronger correlations between features in the same group, compared to features in different groups. For instance, the Botometer and BERT features appear to have moderately strong correlations for features within the same group. As we have previously discussed, this is of no surprise for the Botometer features because we would expect 'bot-like' accounts to display similar characteristics, regardless of the type of bot. Whereas the high within-group average absolute correlation for the BERT features is slightly more surprising. This is because the BERT features are optimised to display different contextual information about language, so we would expect them to exhibit weak correlations. However, the moderately strong inter-group correlations for these features may be due to the nature of the problem which BERTweet was pre-trained on. Namely, BERTweet was pre-trained on language modelling and next sentence prediction, rather than personality profiling. So it is likely the BERT features in our model are less informative for our specific language task. As a result, we found it appropriate to perform a Principal Component Analysis (PCA) on the features prior to including them in our

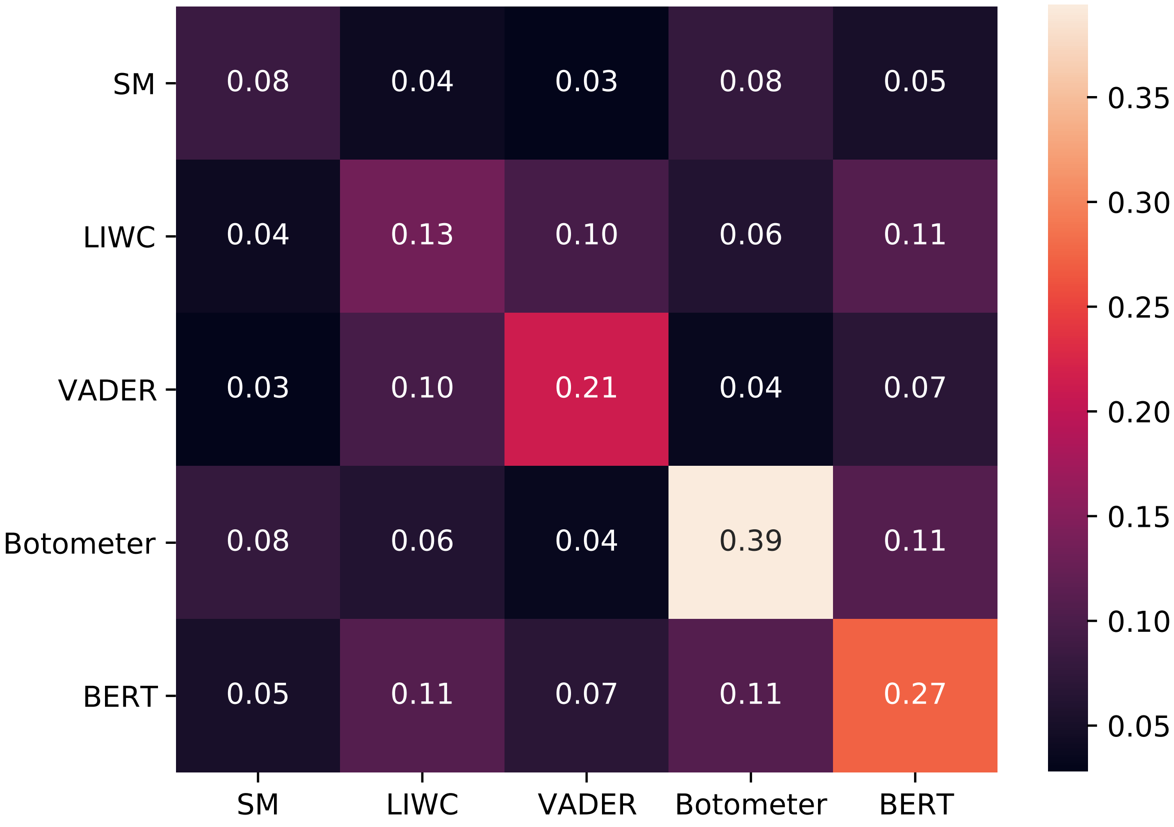


Figure 3.5: Average absolute Pearson correlations within and between different groups of features, calculated using Equation 3.1. In our calculations, we do not include correlations between the same individual features because it will always be one. We take the absolute value of pairwise correlations between the features to avoid them negating in the sum. Each panel displays the average absolute correlation value and can be interpreted as the average correlation strength within and between different groups of features.

model – an overview of PCA is provided in Section 2.4.3. In Figure 3.6, we provide the cumulative percentage of explained variance for each of the principal components, where we order the principal components based on their explained variance of the data. Note that the data was standardized prior to performing the PCA.

In Figure 3.6, we observe almost all of the variance to be explained by the first 200 principal components. In particular, the first principal components explains 25.1% of the variance in the data and the first 200 principal components explain 95.4% of the variance in the data. Note, we include the black vertical and horizontal lines on the figure to demonstrate the amount of variance explained by the first 200 principal components. As a

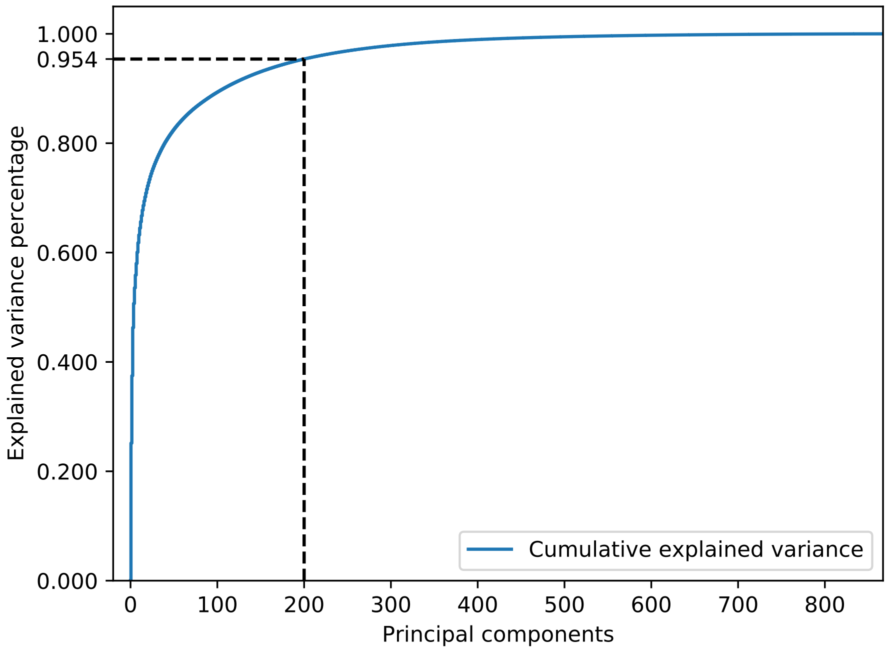


Figure 3.6: Cumulative explained variance of the principal components calculated on the features in our dataset. The explained variance of each principal component is calculated based on their eigenvalues - as discussed in Section 2.4.3. We include the horizontal and vertical black dotted lines to indicate the explained variance of the first 200 principal components – this is also the number of principal components we include in our subsequent models. Note that the features were standardized prior to performing the PCA.

result, we decided to utilise the first 200 PCA components in our machine learning models, rather than the raw 866 features in our original data. We did this because a reduction in 4.6% of variance is a trade-off we are prepared to make in order to reduce the dimension of the feature space by 666 dimensions. This not only improved the computational time of fitting our models, but also significantly reduced the multicollinearity between the features we include in our models. This is of particular importantance for the logistic regression models, where we require the features to not be dependent on each other.

The work in this section has allowed us to better understand our dataset. We have;

- Understood the balance of Myers-Briggs types in our dataset and how sampling procedures may have biased our dataset.

- Discussed the imbalance of the dichotomies in our dataset and how these imbalances may arise.

- Analysed the independence of the Myers-Briggs dichotomies and how this will impact the architecture of our models.
- Compared the correlations and multicollinearity of different features in the dataset within and between different feature groups.
- Performed a principal component analysis on the features in our dataset due to concerns with the size of our feature space and the multicollinearity of different features

This analysis of our dataset will fundamentally shape how we model the personality types of individuals on Twitter. In the next section we train a number of different machine learning models on the dataset and compare the performance of these models.

3.2 Personality Profiling Models

In this section, we develop and compare a number of different machine learning models in predicting the Myers-Briggs personality type of users on Twitter. To do this, we utilise the labelled dataset we have discussed in Section 3.1. In each case, we perform four independent models on each of the four Myers-Briggs dichotomies. Firstly, we compare a number of different techniques for dealing with class-imbalance problems by assessing the performance of these techniques on a logistic regression classifier. We then compare the results of this logistic regression classifier to more sophisticated machine learning models, like naive Bayes, support vector machines and random forests. In each model, we assign the first personality trait of each dichotomy to be the 'positive class' and the second personality trait to be the 'negative class' – this is of particular importance when we consider metrics like precision and recall. Finally, we develop a statistical framework for assessing the importance of different features in the models. We discuss how each of these features contribute to someone's personality type and analyse the importance of different feature groups in the model.

3.2.1 Logistic Regression Classifier

First we perform a logistic regression classifier on each of the four dichotomies in our dataset – an overview of the logistic regression model is provided in Section 2.4.1. We perform independent classifiers on each of the dichotomies because we observe low values for the Cramér's V Statistic in Figure 3.3, indicating there is very little dependency between each of the dichotomies. Moreover, the class imbalances we observe for some of the dichotomies (particularly Intuitive/Sensory and Extrovert/Introvert), leads us to perform different weighting/sampling techniques on our dataset prior to model fitting. Namely, we compare the results of a standard logistic regression model (with no sampling techniques) to five different schemes for dealing with imbalance in the response variable. These class imbalance techniques include models where we:

- Weight the importance of classifying dichotomies,
- Upsample the minority class (with replacement),
- Perform the Synthetic Minority Oversampling Technique (SMOTE), as described in Section 2.4.3, on the minority class,
- Downsample the majority class, and
- Perform a combination of downsampling the majority class and SMOTE on the minority class.

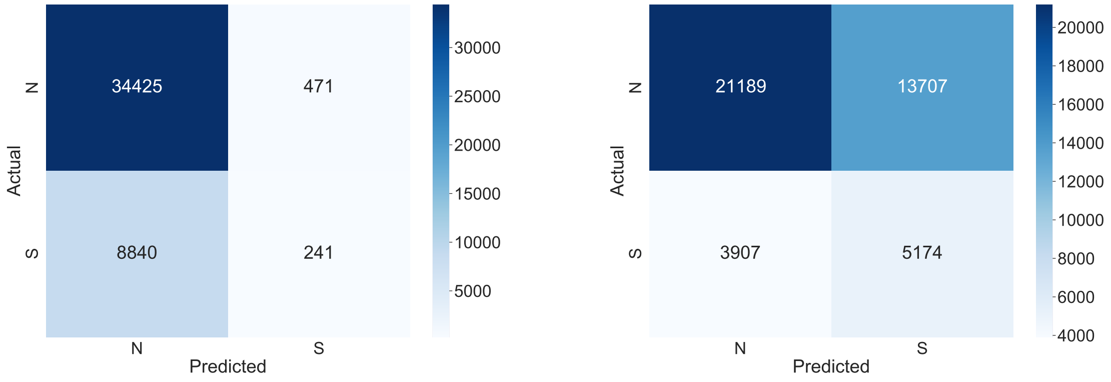
The weighted model we performed involved altering the log-likelihood of the logistic regression model in Section 2.4.1. In this case, we weight each component of the log-likelihood by the inverse frequency of the class size. Hence, the log-likelihood becomes;

$$\ell = \sum_{i=1}^{n} \left[\frac{Y_i}{w_1} \log\left(p\left(Y_i = 1 \mid \mathbf{x}_i\right)\right) + \frac{1 - Y_i}{w_0} \log\left(p\left(Y_i = 0 \mid \mathbf{x}_i\right)\right) \right],$$

where $w_0 = \sum_{i=1}^{n} I_{\{Y_i=0\}}$ and $w_1 = \sum_{i=1}^{n} I_{\{Y_i=1\}}$ are the weights which we set equal to the observed number of subjects in each class. Note that we define all other variables as we did in Section 2.4.1. In the upsampled, SMOTE and downsampled models, we perform the sampling methods until the class balance of each dichotomy is equal in the training set. Note that the upsampled and SMOTE techniques will result in both classes being equal in size to the majority class, and the downsampled technique will result in both classes being equal in size to the minority class. For the model involving a combination of SMOTE and downsampling, we combined the techniques so that the number of sampled in the minority class was at least 80% of the number of samples in the majority class. In particular, we sampled from the classes to ensure the ratio of samples in the minority class was 90%, 80%, 85% and 100% of the samples in the majority classes for the Extroverted/Introverted, Intuitive/Sensory, Thinking/Feeling and Judging/Perceiving dichotomies, respectively. Note that we only perform these sampling techniques on the training set prior to fitting every model – we do not alter the testing set as it is important to reflect the true class proportions in this set in order to emulate our models performance on realistic data.

Model Accuracy

For each model, we used the first 200 principal components of the features described in Tables 3.2 and 3.4 as our predictors. These principal components were calculated using a PCA on the training data – a mathematical explanation of PCA is provided in Section 2.4.3. Using the principal components enabled us to reduce the multicollinearity between the features and also reduced the computational time when fitting the model. In each case, we perform cross validation with ten splits on the dataset. In Figure 3.7 we include the confusion matrices for the Intuitive/Sensory dichotomy under the standard model and the upsampled model.



(a) Standard logistic regression

(b) Upsampled logistic regression

Figure 3.7: Confusion matrices when applying the standard logistic regression model (left) and the upsampled logistic regression model (right) to the Intuitive/Sensory dichotomy. We include the results from the Intuitive/Sensory dichotomy because it is the most unbalanced dichotomy and best demonstrates how a sampling technique impacts the predictions from a model.

In Figure 3.7, we can observe that the standard logistic regression model is primarily predicting the majority class, with very few predictions for the minority class. This indicates that this model is predominately exploiting the class imbalance to make predictions on the testing sets. In comparison, the upsampled model predicts significantly more of the minority class on the testing sets – resulting in more accurate predictions for the minority class. This highlights the importance of using these various weighting/sampling techniques; to reduce the extent to which a model exploits class imbalances in its predictions. Because of this, these models will utilise more information contained in the features about someone's personality type, resulting in models which are more informative of the data. However, in Figure 3.7 we also observe an increase in the number of majority class (Intuitive) individuals who we predict to be in the minority class (Sensory). Consequently, we are observing a clear trade-off between accurately predicting the majority class and accurately predicting the minority class. So we expect that using these various weighting/sampling techniques will reduce the overall accuracy of our models – simply because there are more observations in the majority class than the minority class. In order to demonstrate this, we firstly report the accuracies of the various logistic regression models in Table 3.5. We report four types of accuracy depending on the number of accurately predicted dichotomies in each model. These are the proportion of users where we accurately predict all four dichotomies, at least 3 dichotomies, at least 2 dichotomies and at least one dichotomy. We acknowledge that accuracy can be misleading metric when assessing a models performance on unbalanced data. Hence, we report the accuracies for a random classifier and a majority class classifier. The random classifier demonstrates the accuracy if we were to predict a random personality type for each user and the majority class classifier demonstrates the accuracy if we were to predict the majority class for every user. We report the theoretical accuracies for the random classifier and we determine the accuracies for the majority classifier using the proportion of types in Figure 3.1. The majority class classifier is included to indicate how informative the features are in our models compared to the class imbalances in the dataset. This enables us to report these accuracies without misleading readers into believing our results are better than they are.

	Accurately Predicted Dichotomies			
Model	4	≥ 3	≥ 2	≥ 1
Standard	**20.82**	**60.43**	**89.35**	**98.82**
Weighted	13.78	48.89	82.84	97.69
Upsampled	13.93	48.77	82.79	97.69
SMOTE	13.89	48.63	82.51	97.65
Downsampled	13.70	48.72	82.54	97.70
SMOTE + Downsample	16.40	53.20	85.11	98.08
Random Classifier	6.250	31.25	68.75	93.75
Majority Class	15.31	54.54	87.20	98.28

Table 3.5: Reported accuracies for the logistic regression models. For each model, we include the accuracy of correctly predicting all four dichotomies, at least three dichotomies, at least two dichotomies and at least one dichotomy. We report the theoretical accuracies of a random classifier and we determine the accuracies of a majority class classifier using the proportion of types observed in Figure 3.1.

The results in Table 3.5 indicate that the standard logistic regression model achieved a better accuracy than all other models, including the random classifier and majority class classifier. Furthermore, we observe that the SMOTE + downsampled model outperforms the majority class classifier at accurately predicting all four dichotomies, but not in any of the other criteria included in Table 3.5. For all other models, we observe they do not succeed in outperforming the majority class classifier. Hence, the techniques we use to fix issues with class imbalance result in models with a lower accuracy, as we previously hypothesised. However, the purpose of these models was not to achieve a very high accuracy, but rather to accurately predict more minority personality types – which we previously observed in Figure 3.7. In Table 3.5, we can observe a clear trade-off between achieving a high accuracy and how much we balance the classes. For instance, we see that the standard model (with the most imbalance) achieves the highest accuracy and the models which aim to completely balance (or weight) the classes achieve the lowest accuracies. Whereas the SMOTE + downsampled model achieves accuracies somewhat in the middle, and this model improves the class imbalance, but does not completely balance the classes. Hence, we observe that the class-imbalances in our data are somewhat improving the accuracy of our models; the standard model is exploiting the class imbalance in our dataset to make predictions. The standard model is clearly also utilising information in the features to make its classifications, further improving its accuracy from what we observe with the majority class classifier. However, this improvement in accuracy only exceeds the majority class classifier by 5.51% – which is only a marginal increase. As we have previously discussed, accuracy alone is not a great metric for analysing how well our models perform on unbalanced datasets. This is because models can exploit these class imbalances when performing predictions, leading to accuracies which are not informative of a model's performance. Hence, it is important to consider Receiver Operating Characteristic (ROC) curves as well as metrics like Area Under the Curve (AUC) when assessing the performance of our models on unseen data.

Receiver Operating Characteristic Curves

Next we considered the Receiver Operating Characteristic (ROC) curve of our models performance on testing data. ROC curves provide a good representation of a models performance on our dataset and are less sensitive to class-imbalances in the data compared to other metrics like accuracy. Training the models independently for each dichotomy means that there is a total of six models of the four different dichotomies, a total of 24 different logistic regression models. As a result, creating one ROC curve for each model would make visualisations a lot clearer. An approach used by other authors such as Gunčar *et al.* [65], De *et al.* [39] is to macro-average and micro-average the true positive rate and false positive rate of points on the ROC curve for each model (or dichotomy in our case). Consequently, the macro-average True Positive Rate (TPR) for some probability threshold, α, is given by:

$$\frac{\mathrm{TPR}_\alpha^{\mathrm{E/I}} + \mathrm{TPR}_\alpha^{\mathrm{N/S}} + \mathrm{TPR}_\alpha^{\mathrm{T/F}} + \mathrm{TPR}_\alpha^{\mathrm{J/P}}}{4}, \tag{3.2}$$

where TPR_α^x is the true positive rate of model, x, with threshold, α. Note that the macro-average False Positive Rate (FPR) is also an average of the four FPRs for each dichotomy. The micro-average TPR for some probability threshold, α, is given by:

$$\frac{TP_\alpha^{\mathrm{E/I}} + TP_\alpha^{\mathrm{N/S}} + TP_\alpha^{\mathrm{T/F}} + TP_\alpha^{\mathrm{J/P}}}{TP_\alpha^{\mathrm{E/I}} + TP_\alpha^{\mathrm{N/S}} + TP_\alpha^{\mathrm{T/F}} + TP_\alpha^{\mathrm{J/P}} + FN_\alpha^{\mathrm{E/I}} + FN_\alpha^{\mathrm{N/S}} + FN_\alpha^{\mathrm{T/F}} + FN_\alpha^{\mathrm{J/P}}}, \tag{3.3}$$

where TP_{α}^{x} and FN_{α}^{x} denote the number of true positives and false negatives in the data used for model, x, with threshold, α. The micro-average FPR is calculated in the same way, except with TP replaced with the False Positive (FP) and FN replace with True Negative (TN). We performed these calculations at each threshold for every logistic regression model type and displayed the results in an ROC curve, seen in Figure 3.8. We also include the Area Under the Curve (AUC) metric for each of the ROC curves.

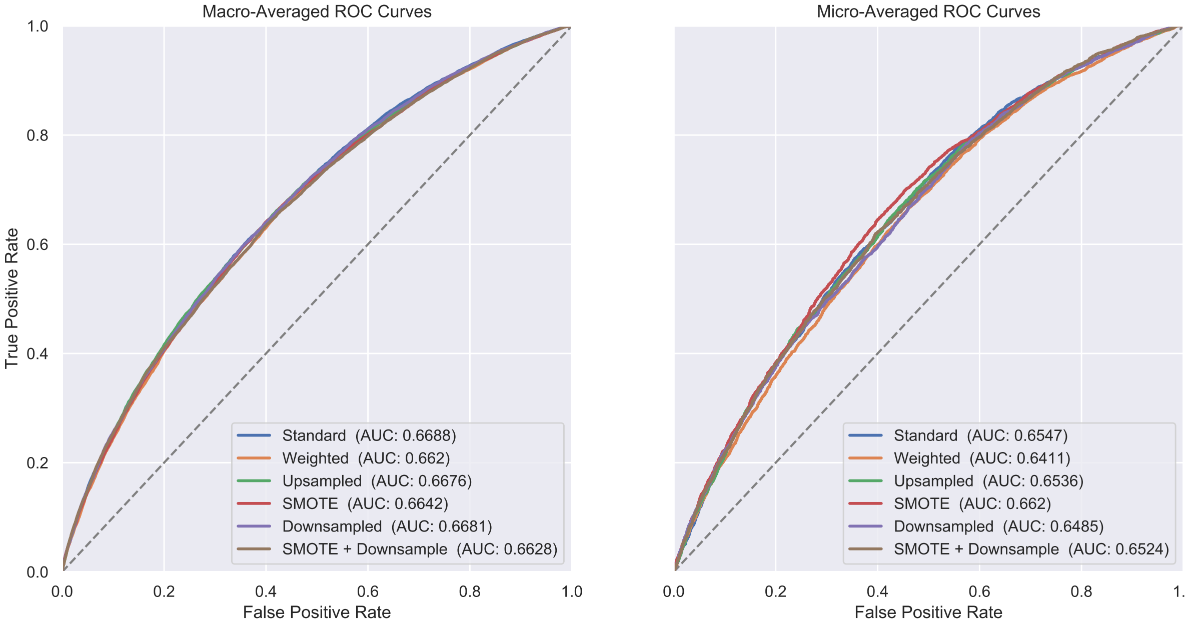


Figure 3.8: Macro-averaged (left) and Micro-averaged (right) Receiver Operating Characteristic (ROC) curves for each logistic regression model. Note that we macro-average and micro-average the results from the independent models for each dichotomy using Equations 3.2 and 3.3, respectively. We include the Area Under the Curve (AUC) metric for each of the ROC curves as well.

Figure 3.8 show similar performance for all models (based on the AUC). In more detail, the best performing macro-averaged model is the standard model. Whereas, the best performing micro-averaged model is the SMOTE model. We can interpret the macro-average ROC curve as a measure of average performance for each of the four dichotomies, where each models TPR or FPR is weighted equally. Whereas, the micro-average ROC curve aggregates the contributions of all samples in each model and weights individual predictions equally. Hence, models with few positive samples will have a smaller contribution in the micro-averaged TPR compared to the macro-averaged TPR. Similar is true for the FPR, where models with a lower number of FPs will have a smaller contribution to the micro-averaged FPR, compared to the macro-averaged FPR. As a result, we can use the macro and micro-averaged results to determine how each model performs across dichotomies and individuals, respectively. Moreover, the micro-averaged AUC will generally be less sensitive to class imbalances in the data, making it more comparable across models. Using this information, we can deduce that the standard model performs best when averaging across each dichotomy, whereas the SMOTE model performs better on an individual level. We also observe that the macro-averaged AUCs are consistently larger than the micro-averaged AUCs. This indicates that all of the models perform better across dichotomies, rather than on an individual level.

Performance of Binary Classifiers

Next we considered the performance of binary classifiers for each independent dichotomy. We calculate a number of metrics for each model on the results from the cross-validation with ten splits. These metrics include the AUC, F1, Precision, Recall and Accuracy, and are displayed in Figure 3.9. We include these metrics to demonstrate our model's performance on unseen data. However, it is important to note that the accuracy, precision, recall and F1 score of the models are sensitive to class imbalance in the training set. Whereas, the AUC score is less sensitive to class imbalance in the training set – this is because AUC equally weights the true positive rate and false positive rate. Consequently, AUC provides a more robust metric for measuring the performance of our models on unbalanced data.

In Figure 3.9, we observe very little changes in the metrics for the Judging/Perceiving dichotomy – this is due to the data for this dichotomy already being fairly balanced (see Figure 3.2). Whereas, the Extrovert/Introvert, Intuitive/Sensory and Thinking/Feeling dichotomies are all sensitive to adopting techniques for fixing issues with class imbalance. Of these dichotomies, we observe very little difference in the metrics for the upsampled, downsampled, weighted and SMOTE models. However, there is a noticeable difference in the metrics for the SMOTE + downsampled model and a clear difference in the metrics for the standard model. The standard model performs better in terms of accuracy but worse for other metrics like F1 and either Precision or Recall. The metrics for the SMOTE + downsampled model lie in the middle of the standard models results and the results of all other models. As we have previously mentioned, the first personality trait of each dichotomy is assigned to be the positive class for each model. So we observe different behaviour for the precision and recall of a model depending on whether the first trait or second trait of each dichotomy is the majority class. Typically, we only observe this behaviour in the standard and SMOTE + downsampled models because we generally do not equate class sizes under each dichotomy for these models. For instance, we observe a lower recall and larger precision for models where the negative class is the majority class, such as the Extroverted/Introverted model or the Thinking/Feeling model. Whereas, we observe a larger recall and lower precision for models where the positive class is the majority class, such as the Intuitive/Sensory model. We also observe these effects to be

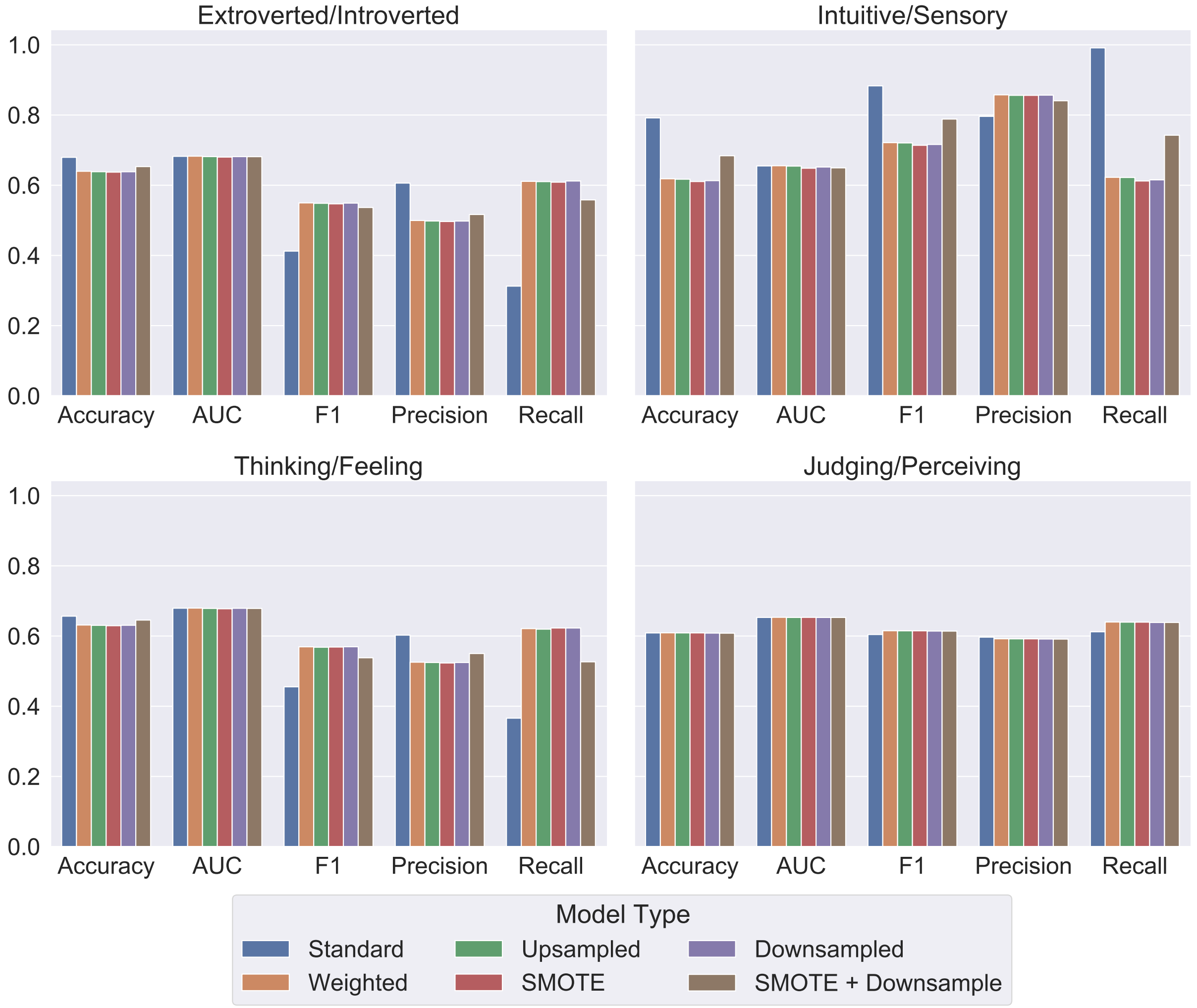


Figure 3.9: Summary of metrics from the logistic regression models of the four independent Myers-Briggs dichotomies. These metrics include the Accuracy, Area Under the Curve (AUC), F1, Precision and Recall Scores. For each model, these metrics were calculated based on a ten-fold cross validation

more pronounced for the Intuitive/Sensory model, where we obtain nearly a perfect recall for the standard model – this is because of the severity of the class imbalance for this dichotomy.

More generally, the Accuracy and AUC of each model gives us a good indication of their predictive capability and their performance on our data, respectively. We can observe that most of the accuracies and AUCs lie around 0.65 (65%) with the odd exception for modelling severely imbalanced dichotomies with the standard model. While 65%

is a reasonable accuracy, it lies below a number of the accuracies achieved by other authors. For instance, Bharadwaj *et al.* [15] achieves accuracies of up to 90% on data from `www.personalitycafe.com` – more results from Myers-Briggs personality profiling models by other scholars can be observed in Table 2.3. However, as we have previously discussed; a number of these authors did not acknowledge any of the class imbalance problems which arise in these datasets and consequently their accuracies may be misrepresentative of their models performance on unseen data. We also acknowledge that our dataset has been sourced using different methods to other datasets and may resultantly be harder to model. Plank and Hovy [117] obtained the only other Twitter dataset of labelled MBTI's (to our knowledge) and hence the results from their study provide a good baseline to compare our results to. However, Plank and Hovy [117] are one of many scholars in this area which do not report the AUC metric – so we are constrained to comparing our models using the accuracies. Moreover, these scholars do not address problems which arise due to class imbalance and this can be consequently seen by viewing their confusion matrices (where some personality types are never predicted). As a result, in Table 3.6 we compare the results for our standard logistic regression model to the results obtained by Plank and Hovy [117] on Twitter data and Bharadwaj *et al.* [15] on the Personality Cafe data.

Dichotomy	Our Accuracy (Twitter)	Plank/Hovy Accuracy (Twitter)	Bharadwaj Accuracy (Personality Cafe)
E/I	68.0	**72.5**	85.3
N/S	79.2	**79.5**	90.4
T/F	**65.7**	61.2	87.0
J/P	**60.9**	58.2	79.0

Table 3.6: Comparison between the accuracies obtained from our standard logistic regression model and the accuracies from models by Plank and Hovy [117] and Bharadwaj *et al.* [15]. Plank and Hovy [117] trained a logistic regression model on a dataset from Twitter using n-grams and SM features – their model use no weighting/sampling techniques to deal for class imbalances. Bharadwaj *et al.* [15] trained a number of machine learning/deep learning models on the Personality Cafe Data using TF-IDF and LIWC features – these authors also did not address class imbalances. Note that we display the best performing Twitter model for each dichotomy in bold.

The results in Table 3.6 demonstrate that our accuracies are very similar to those obtained by Plank and Hovy [117]. In particular, we observe that our Thinking/Feeling and Judging/Perceiving models outperform the Plank and Hovy [117] models, however the Plank and Hovy [117] Extroverted/Introverted and Intuitive/Sensory models outperform ours. Moreover, we observe that the Bharadwaj *et al.* [15] models on the My Personality data outperform all of our models for the accuracy metric. This suggests that an account's

MBTI type is more difficult to model Twitter compared to other social media platforms like Facebook, Reddit and the Personality Cafe – potentially because there is more text data on these other platforms. It is also significant that we have biased our dataset by querying for users who self-report their type, and these accounts may be more difficult to model than accounts which do not self-report their type. A number of other authors have opted for different techniques of sourcing labelled data, such as, Golbeck *et al.* [60] and Howlader *et al.* [71] who sourced their data by administering an API providing an online questionnaire. While online questionnaires are expensive, we would expect them to provide better results and a smaller misclassification rate. While our data sampling techniques were novel and produced the largest corpus of labelled Twitter accounts with a MBTI, they also lead to a small number of accounts being mislabelled (as seen previously in Table 3.1). Hence, our results will also be reflective of the misclassified accounts and these would further reduce the performance of our models.

In this section we considered the performance of a logistic regression classifier for predicting the personality types of users in our dataset. We observed that performing various weighting/sampling techniques improved the performance of our models on minority classes but reduced the accuracy of these models overall. We found the most accurate classifier to be the standard logistic regression classifier, however this model only achieved an accuracy 5.51% larger than a majority class classifier. Hence, we deduce that the logistic regression model is not a great fit for our data. So we will explore a suite of different models in the subsequent sections including a naive Bayes classifier, support vector machines and random forests.

3.2.2 Model Comparison

In this section we will compare the results of several models used to predict the Myers-Briggs personality type of users in our dataset. In particular, we compare the results of a naive Bayes classifier, support vector machine classifier (SVM) and random forest classifier to the aforementioned results for the logistic regression classifier. We implemented the naive Bayes, support vector machine and random forest classifiers on each of the four dichotomies in the personality dataset – an overview of these classifiers is provided in Section 2.4.1. For each model, we build four independent classifiers on each dichotomy. Note that the choice to build independent classifiers has previously been justified using the results from the Cramér's V Test (see Figure 3.3). Similar to the logistic regression models, we perform different weighting/sampling techniques on our data due to the class imbalances. We use the same sampling techniques as we developed in Section 3.2.1, however each weighted model is performed slightly differently to the weighted logistic regression model. In what follows, we will explain the approach used to weight each of the models.

For the weighted naive Bayes model, we incorporate a weighting factor into the class priors which aims to balance predictions across the classes. In particular, we alter the class priors in Section 2.4.1 so they are both equiprobable:

$$P\left(Y_i = 0\right) = P\left(Y_i = 1\right) = 0.5, \quad i = 1, \ldots, n,$$

where we define all other variables as in Section 2.4.1. The likelihood and scaling factor for the naive Bayes model is then calculated using the new weighted priors.

For the weighted support vector machine model, we include weighting factors into the inverse regularization parameter for each class. Recall in Section 2.4.1 we defined C as the inverse regularization parameter which controls for the strength of the penalty associated with the distances of our data, ζ_i, from the fitted hyperplane. Note that larger values for C will mean the penalty for the ζ_i's is weighted more and so the model will have a preference to correctly classify data with larger values of the inverse regularisation parameter. Hence for the weighted SVM model, we weight the inverse regularisation parameter so it is inversely proportional to the class frequency of each class:

$$C_0 = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{\{Y_i=0\}}} \quad \& \quad C_1 = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{\{Y_i=1\}}}.$$

The primal problem in Section 2.4.1 then becomes:

$$\begin{array}{ll} \min\limits_{w,b,\zeta} & \frac{1}{2} w^T w + C_0 \sum_{i:Y_i==0} \zeta_i + C_1 \sum_{i:Y_i==1} \zeta_i, \\ \text{subject to} & Y_i \left(w^T \phi\left(\mathbf{x}_i\right) + b\right) \geq 1 - \zeta_i, \\ & \zeta_i \geq 0, \quad i = 1, \ldots, n, \end{array}$$

where we define all other variables as we did in Section 2.4.1.

For the weighted random forest model, we incorporate weights into the class probabilities for each candidate split. This means the model will preference accurately predicting samples in the minority class over the majority class. We alter the class probabilities in Section 2.4.1 such that they are proportional to the inverse class frequencies:

$$P\left(Y_i = y \mid \theta\right) \propto \frac{1}{n_m \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=y\}}} \sum_{Y_i \in Q_m} \mathbb{1}_{\{Y_i=y\}}, \quad y \in \{0, 1\},$$

where we define all other variables as we did in Section 2.4.1. We then renormalise these probabilities so that they sum to one and calculate the Gini Impurity loss function using the new renormalised class probabilities:

$$H\left(Q_m \mid \theta\right) = 2P\left(Y_i = 0 \mid \theta\right) P\left(Y_i = 1 \mid \theta\right).$$

We then minimise the overall impurity function for each candidate split, as described in Section 2.4.1.

In addition to the weighted models, we also performed upsampling, SMOTE, downsampling and a combination of SMOTE/downsampling on the naive Bayes, support vector machine and random forest models. We performed the upsampling, SMOTE and downsampling until the class balance of each dichotomy in the training set is equal. For the models combining SMOTE and downsampling, we sampled the classes so they had the same ratios of observations as the the 'SMOTE + Downsampled' logistic regression models. We used the same ratios as the logistic regression models for consistency – this also allowed us to compare the results between the different models. Note that we only performed these sampling techniques on the training set in each case. Moreover, we used the first 200 principal components as predictors in each models – again, these were determined using only the training data. While the naive Bayes classifier was very quick to train and use during testing, reducing the dimension of the predictor space significantly improved the training time for the support vector machine and random forest classifiers. The computational time of fitting the support vector machine Models in our analysis scale between $\mathcal{O}\left(n_{\text{features}} \times n_{\text{samples}}^2\right)$ and $\mathcal{O}\left(n_{\text{features}} \times n_{\text{samples}}^3\right)$. Since removing training data was not an option, reducing the dimensionality of the feature space was an obvious choice, and significantly reduced the computational time when training the support vector machine classifiers. Moreover, the computational time of fitting the random forest classifiers in our analysis is $\mathcal{O}\left(n_{\text{features}} \times n_{\text{samples}} \times \log n_{\text{samples}}\right)$. Again, removing training data was not an option for the same reasons as above, so reducing the dimensionality of the feature space was sensible. This also significantly reduced the computational time when training the random forest classifiers.

For each model, we performed 10-fold cross validation. We include all the results in Appendix A, where the naive Bayes, support vector machine and random forest results are displayed in Sections A.1, A.2 and A.3, respectively. We include the summary of metrics (Accuracy, AUC, F1, Precision and Recall) for the naive Bayes, support vector machine and random forest models in Figures A.1, A.3 and A.5, respectively. We use these results to compare the performance of the different models, along with the summary of metrics for the logistic regression models (provided in Figure 3.9). The first thing to observe from these figures is how badly the naive Bayes classifier performs on this dataset. We observe significantly lower accuracies, AUCs and F1 scores for the naive Bayes model, and in particular, the two naive Bayes models which utilise SMOTE. The poor performance of the naive Bayes classifier is best summarised by viewing the accuracies in Table A.1. In this table, we see that the majority class classifier outperforms all of the various naive Bayes classifiers which we explore. Adding to this, we observe the two naive Bayes classifiers utilising SMOTE to have a significantly worse performance than any of the other naive Bayes classifiers, and in some cases worse then the random classifier. This indicates there may be a flaw when combining SMOTE with naive Bayes, something which is often overlooked by other researchers. We explore the reasoning for why SMOTE performs poorly with naive Bayes in Section 3.2.3 below.

Across the four different types of machine learning models, we observe the support vector machine models perform the best, on average (across all metrics). Table 3.7 com-

	Accurately Predicted Dichotomies				AUCs	
Model	4	≥ 3	≥ 2	≥ 1	Macro	Micro
Standard LR	20.82	**60.43**	89.35	98.82	0.6688	0.6547
SMOTE LR	13.89	48.63	82.51	97.65	0.6642	**0.6620**
Standard NB	14.20	49.17	81.91	97.40	0.5784	0.5867
Upsampled NB	13.75	48.06	80.82	97.18	0.5861	0.5917
Standard SVM	**20.95**	60.25	**89.64**	**98.90**	**0.6693**	0.6518
SMOTE SVM	13.56	48.61	82.54	97.61	0.6660	0.6554
Standard RF	19.69	57.96	88.69	98.67	0.6223	0.6273
Upsampled RF	19.70	58.16	88.48	98.76	0.6305	0.6264
Random Classifier	6.250	31.25	68.75	93.75	0.5000	0.5000
Majority Class	15.31	54.54	87.20	98.28	0.5000	0.5000

Table 3.7: Reported accuracies and AUCs for the best performing models from each Machine Learning model. For each machine learning model, we include the results from the 'Standard' model (where we use no weighting/sampling) as well as the results from the best performing weighted/sampling model (excluding the 'SMOTE + Downsampled' model). We exclude the 'SMOTE + Downsampled' model because it does not completely address class imbalances, and it would resultantly not be a fair comparison. Note that we determine the 'best performing weighted/sampling model' based on the sum of their macro-averaged and micro-averaged AUCs. For each model, we include the accuracy of correctly predicting all four dichotomies, at least three dichotomies, at least two dichotomies and at least one dichotomy. We report the theoretical accuracies of a random classifier and we determine the accuracies of the majority class classifier.

pares the accuracies and AUCs of the best performing models from each machine learning model. In each case, we include the 'Standard' model and the weighted/sampling model which achieves the highest sum of the micro-averaged and macro-averaged AUC scores. We see an obvious trade-off between training time and accuracy. For instance, the poorest performing model is the naive Bayes model, which is very fast to train, whereas the support vector machine model performs the best, and is the longest to train. The results for the logistic regression and the random forest models lie between the SVM and naive Bayes models. However, the standard logistic regression model most accurately predicts at least three out of four of a user's dichotomies and performs only marginally worse than the standard SVM model for all other metrics. This indicates that while logistic regression models are relatively simple, they can be highly accurate models, even when compared to more complex models like support vector machines and random forests. Adding to this, the logistic regression model is also significantly faster to train than the support vector machines – making it the model of choice if computational time is of concern or if we were training the model on larger datasets. However, since these models only need to be trained once, the support vector machine model would still be the preferred model for our current dataset. This is validated by the results in Table 3.7 which show that the standard SVM model performs the best when average across the provided metrics.

Moreover, SMOTE is the best performing weighting/sampling technique for the logistic regression and support vector machine models, and upsampling is the best weighting/sampling technique for the naive Bayes and random forest models. We deduce that the SMOTE and upsampling techniques appear to be the best performing weighting/sampling methods from the selected techniques. We believe that downsampling does not perform as well as these techniques because downsampling essentially involves removing data from our training set, reducing the information provided to our models. It is more difficult to justify why the weighted models perform worse than the two oversampling methods; perhaps using weights inversely proportional to class frequencies puts too much weighting on the minority classes.

3.2.3 Drawbacks of using SMOTE with Naive Bayes

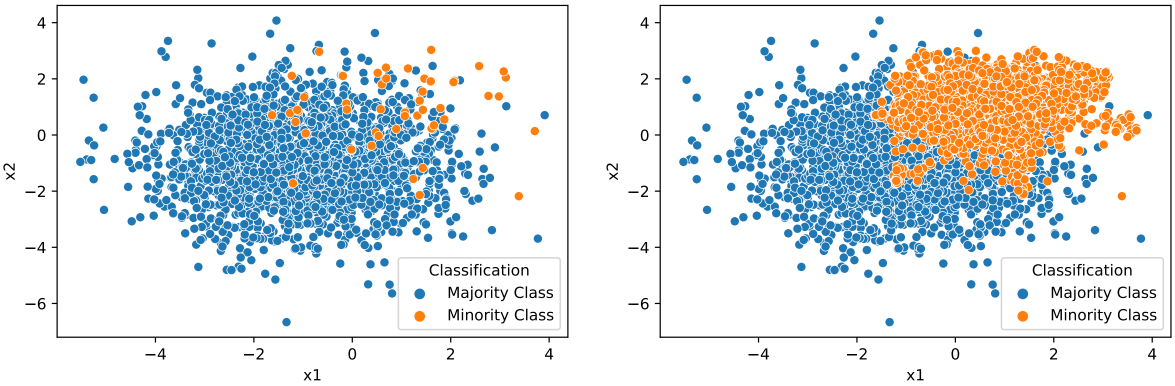
In this section we will explore the drawbacks of combining SMOTE with naive Bayes on unbalanced data, where there is little separation between the classes. We will demonstrate this in two-dimensions as the results can be easily visualised, however our conclusions can be extended to features of any dimension.

Firstly, we simulated some independently and identically normally distributed data in two dimensions. Let $\mathbf{x}^{\text{maj}}$ and $\mathbf{x}^{\text{min}}$ denote vectors of samples for the majority and minority classes, respectively. We simulate data from the following distributions:

$$x_i^{\text{maj}} \sim \mathcal{N}\left(\begin{bmatrix}-1\\-1\end{bmatrix}, \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}\right) \quad \& \quad x_i^{\text{min}} \sim \mathcal{N}\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}\right),$$

where we sample 2000 observations in the majority class and 50 observations in the minority class. Note that we use a substantial class imbalance to emphasise the effects of SMOTE on naive Bayes – the class imbalance ratios in our personality dataset were provided in Figure 3.2. The unbalanced simulated training data is shown in Figure 3.10a. After sampling our data, we next generated synthetic data for the minority class using SMOTE. This synthetic data was generated until the minority class had 2000 observations, making the two classes balanced. The training data after applying SMOTE is in Figure 3.10b.

Figure 3.10 shows the clear impact of using SMOTE on the minority class. We observe SMOTE to create new synthetic data which is densely populated in the region created by the original samples from the minority class. In particular, if we were to create a convex hull in the original minority class, then every synthetic sample generated using SMOTE would lie within that convex hull. As a result, we hypothesised the distribution of samples



(a) Training data prior to using SMOTE

(b) Training data after using SMOTE

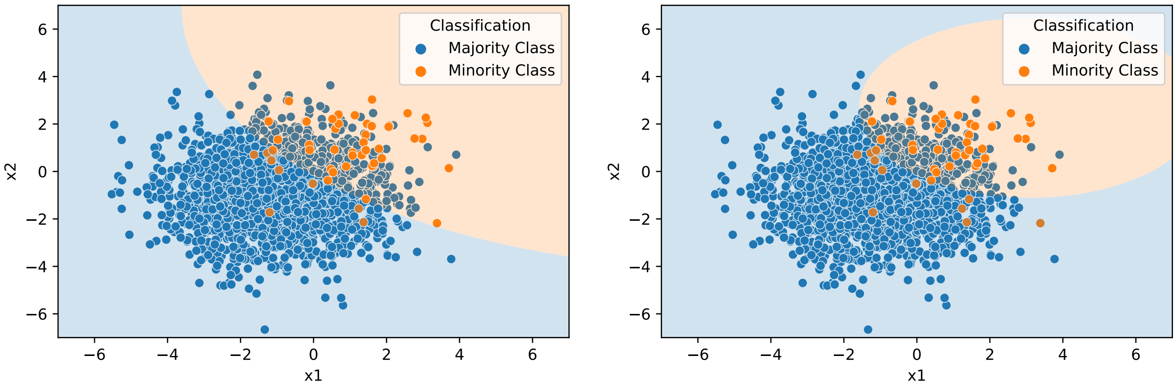
Figure 3.10: Two-dimensional example of a scatterplot of the training data before (left) and after (right) utilising SMOTE. In this case, we use SMOTE to completely balance the two classes. Prior to using SMOTE there were 2000 samples in the majority class and 50 samples in the minority class. An overview of SMOTE is provided in Section 2.4.3.

in the minority class to have a significantly lower variance after performing SMOTE. In this example, we calculated the mean and variance of two variables in the minority class prior to and after performing SMOTE – we display the statistics in Table 3.8.

Variable	Unbalanced	Balanced
$E[x_1]$	0.873	0.867
$E[x_2]$	0.941	1.041
$\mathrm{Var}(x_1)$	1.817	1.156
$\mathrm{Var}(x_2)$	1.531	0.905

Table 3.8: Means and Variances of the two variables on the minority class. We calculate these statistics on the unbalanced and balanced data. The data is balanced using SMOTE.

Table 3.8 shows the variance of the two variables to be significantly lower in the balanced data after applying SMOTE, whereas the means remain relatively similar. This supports our hypothesis that SMOTE will reduce the variance of the features. However, we are interested in the extent to which this will impact the Gaussian naive Bayes classifier. Recall in Gaussian naive Bayes (see Section 2.4.1), we assume the likelihood of the features to have a Gaussian Distribution where we estimate the mean and variance of the features using the sample mean and variance of the training data. As a result, the likelihood of the features in the minority class will have a much narrower Gaussian distribution after applying SMOTE, due to the reduction in variance. We suspect this to in-turn reduce the area of the decision space for samples we predict to be in the minority class. In Figure 3.11, we include the naive Bayes decision boundaries when the model is trained on the unbalanced and balanced data in Figures 3.11a and 3.11b, respectively. Note that in both cases, we include the unbalanced training data in the plots for reference to the decision boundary.



(a) Training data prior to using SMOTE

(b) Training data after using SMOTE

Figure 3.11: The decision boundary from utilising naive Bayes on the unbalanced (left) and balanced (right) training data. In each case, we include the unbalanced training data in the plot for reference to the decision boundary. An overview of naive Bayes is provided in Section 2.4.1.

Consequently, we observe the decision space to be much smaller when the Gaussian naive Bayes model is trained on the balanced data compared to the unbalanced data. As a result, significantly more samples will be classified in the majority class when SMOTE is performed on the unbalanced data. Recall that we perform these weighting/sampling techniques to increase the preference for predictions in the minority classes. Hence, performing SMOTE with naive Bayes has an opposite effect on the predictions of samples in the minority class and we suspect this to be the reason for why naive Bayes performs so poorly when combined with SMOTE on our personality profiling dataset. Our example is only in two-dimensions because results can be nicely visualised. However, these results can be extended to features of any dimensionality, and we expect this problem to be worse in higher dimensions due to the curse of dimensionality. While there are a number of scholars utilising SMOTE and Naive Bayes, such as Flores *et al.* [53] and Saifudin *et al.* [132], this observation has not been reported by any other scholars (to our knowledge).

3.2.4 Feature Importance

In this section, we compare the importance of different features we used to model the Myers-Briggs Personality types of users in our Twitter dataset. To do this, we perform independent upsampled logistic regression models on each of the four dichotomies. As in previous sections, we perform the upsampling for each model until the class balance is equal for each dichotomy. Note that we choose to upsample because it performs well for the logistic regression model (see Section 3.2.1). Upsampled models also do not involve creating 'synthetic' data in the same way that SMOTE does. This is particularly important because we use the t-statistic from each of the parameter estimates to determine the importance of each variable in the regression model – we will explain this in more detail below.

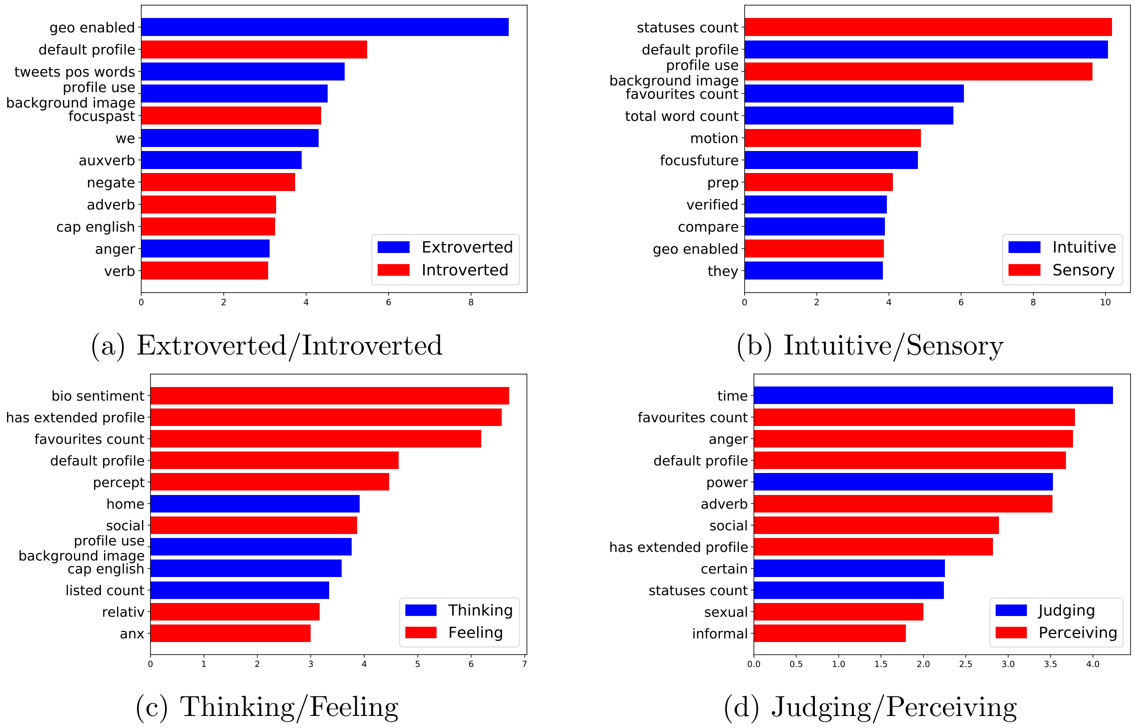
Firstly, we consider the variable importance of the descriptive features in our regression models. In this case, descriptive features include all features except those in the BERT category. For each dichotomy we fit the logistic regression model to the features and perform a stepwise feature selection to obtain a model with only significant features. In each case, we start with a null model and perform the stepwise selection algorithm on the p-values of the features. Moreover, we use a threshold in of 0.05 and a threshold out of 0.1, meaning we iteratively accept features into the model with p-value < 0.05 and remove features from the model with p-value > 0.1. Once there is no more features to add or remove, we calculate the variable importance of different features in the model. We determine the variable importance of the features using the t-statistic for the parameter coefficients associated with each feature. The t-statistic for a parameter estimate of a feature, f, is defined as:

$$t_{\hat{\beta}_f} = \hat{\beta}_f / \mathrm{SE}\left(\hat{\beta}_f\right),$$

where SE $(\cdot)$ denotes the standard error of the parameter. The t-statistic is a good estimate of variable importance because it is a function of the parameter estimate and is scaled by the variance of the parameter estimate. For each dichotomy, we calculate the variable importance of each remaining feature after the stepwise selection algorithm is complete and display the absolute value of the variable importance. Figure 3.12 displays the 12 most important features for each model in a bar chart. We colour the bars based on the variable's preference for each class in the dichotomy. For instance, the most important variable in Extrovert/Introvert model is 'Geo Enabled' and an increase in this variable makes someone more likely to be classified as extroverted, as seen in Figure 3.12a.

Figure 3.12 shows that many features contribute to different dimensions of a Twitter users' Myers-Briggs personality type. Pennebaker and Francis [114] suggested that the words which seem most meaningless sometimes best describe people. These "meaningless" words are called function words and are words such as pronouns (pronoun), personal pronouns (ppron), 1st person singular (i), 1st person plural (we), prepositions (prep), auxiliary verbs (auxverb) and negations (negate). Many of these function words are significant predictors in our models – for example, 1st person plurals are significant in the Extroverted/Introverted model and prepositions are significant in the Intuitive/Sensory model. These findings evidence that researchers need to be mindful of these function words, and performing techniques such as stop-word removal may not be appropriate in many situations.

In particular, we observe that extroverts tend to use more positive language and introverts tend to have more of a focus on the past. This first result is consistent with



(a) Extroverted/Introverted

(b) Intuitive/Sensory

(c) Thinking/Feeling

(d) Judging/Perceiving

Figure 3.12: Variable Importance Plots (VIPs) based on an upsampled logistic regression model for each dichotomy. We determine the variable importance using the t-statistic for the parameters associated with each feature. Variables are sorted by the absolute value of the variable importance (from top to bottom). We colour the bar based on the features preference for each dichotomous class. Note that we do not include BERT features in the VIPs because they're meaning cannot be inferred.

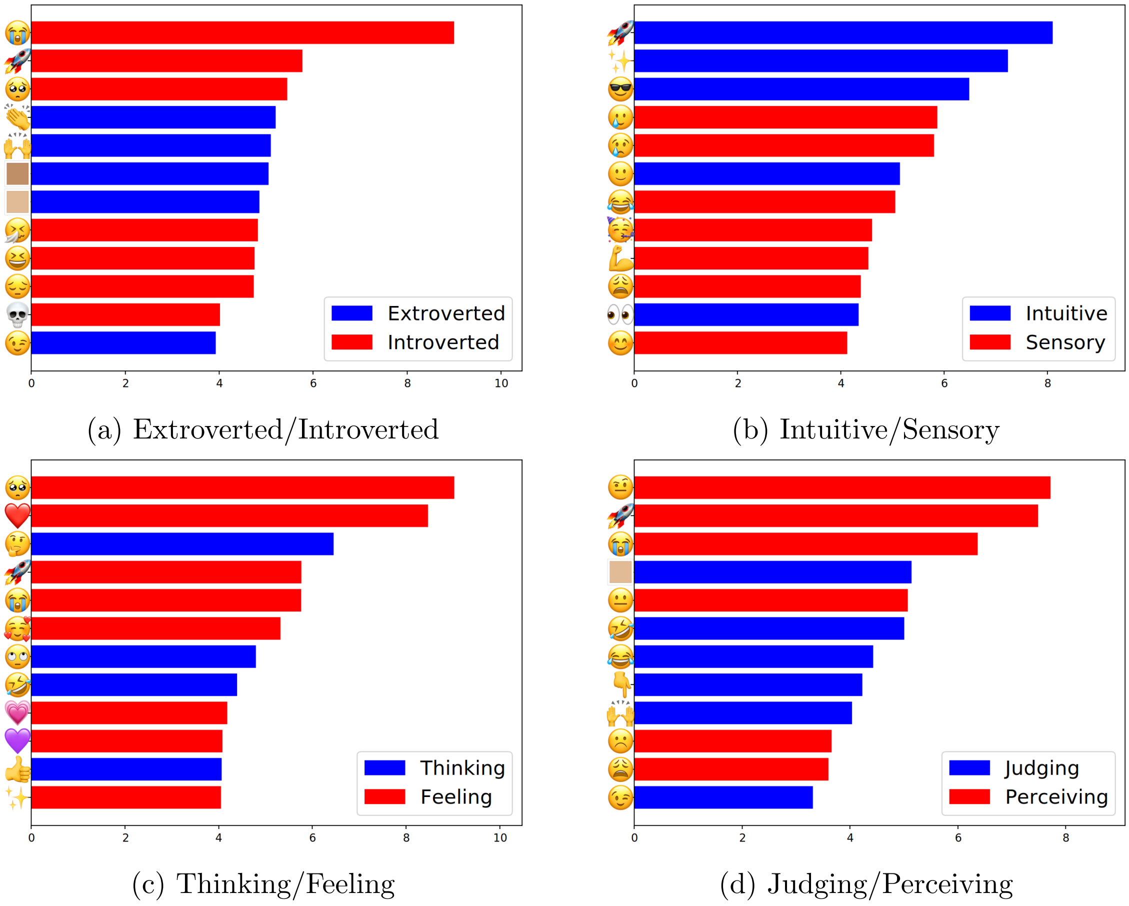
Chen *et al.* [30], who suggested that extroverts display more positive emotion because they have a "dispositional tendency to experience positive emotions". However, we suggest it may also be because extroverts don't use social media as a main form of communication; perhaps extroverts display less positive emotion online because they have a preference to display their negative emotion in offline environments. More controlled psychological experiments would be required to confirm this hypothesis. Introverts are also more likely to be bots. This is consistent with Giorgi *et al.* [58], which showed social spambots tend to be more introverted – however, the effects were not as significant as what we observe here. Perhaps there are other confounding factors which accentuate these effects but aren't included in our models; like age or gender.

Figure 3.12b shows that accounts which have a larger favourites count (i.e. the account likes more tweets) to be more intuitive, whereas accounts which write more statuses tend to be more sensory. We can use the favourites count as a proxy for the amount of information an account consumes, however we acknowledge this proxy is not perfect because people may consume information without liking it. Nonetheless by this proxy, our results suggest that intuitives tend to consume more information on Twitter, whereas sensory individuals tend to write more. Following the definition from The Myers-Briggs Foundation, intuitives pay "most attention to impressions or the meaning and patterns of the information", whereas sensors pay "attention to physical reality, what I see, hear, touch, taste, and smell" [146]. We suggest that we are observing these effects in our results as well; where intuitives are paying attention to the meaning/patterns of various tweets and sensors are writing about what they observe in physical reality. Similarly, Schaubhut *et al.* [136] found that intuitives are significantly more likely to browse, recommend and interact with Twitter content. Larger total word counts make a user more likely to be intuitive, indicating that intuitives write more words per tweet. This shows that while intuitives are less likely to tweet content, when they do so, they tend to write considerably more.

The strongest predictor in the Thinking/Feeling model is the sentiment of the biography, where positive sentiment makes an account more likely to display the 'feelings' characteristic. This indicates that users who display the feelings trait also generally exhibit more positive emotion about themselves, as an account's biography usually reflects information about oneself. Feelings-based accounts also show more angst (anx) and use more social language and perceptual processes (percept) words. Accounts exhibiting the feelings trait have a preference to establish or maintain harmony. Adding to this, we believe that feelings-orientated users tend to use more perceptual process words like 'look', 'heard' and 'feeling' because these users are more in-touch with their emotions. Furthermore, thinking-orientated users tend to be more bot-like – perhaps bot accounts also have a common personality type, which was also observed by Giorgi *et al.* [58].

Finally, the Variable Importance Plot of the Judging/Perceiving dichotomy in Figure 3.12d shows that the strongest predictor is time, where judgers are much more likely to use words related to time and certainty compared to perceivers. 'End', 'until' and 'season' are examples of time-related words and 'always', 'never' are words related to certainty. According to The Myers-Briggs Foundation, judgers tend to "prefer a planned or orderly way of life, like to have things settled and organized" [146]. It appears that judgers use these words because they are concerned about time and want their lives under control as much as possible [100]. Moreover, we observe that perceivers tend to use more social and informal language. This is likely a product of perceivers appearing more "loose and casual than judgers", as suggested by many researchers [146, 100].

Next we explore how emoji usage contributes to the Myers-Briggs Personality Type of Twitter users. On Twitter, emojis often have multiple meanings. For instance, the rainbow flag indicates a user support the user supports the LGBTQ+ social movements, the wave symbolises a "Resister" crowd of anti-Trump Twitter and the okay symbol is used by white supremacists, some of which covertly use the symbol to indicate their support for white nationalism [22]. Hence, we can use these emojis to understand how their meanings interact with different personality types. We did this by determining each emoji's frequency in a user's tweets and included these frequencies as predictors in upsampled logistic regression models. We performed the same stepwise feature selection algorithm as above and display the 12 most important predictors from the remaining models. The results are displayed in Figure 3.13, and again we use the t-statistic from the parameter coefficients to estimate variable importance.



(a) Extroverted/Introverted

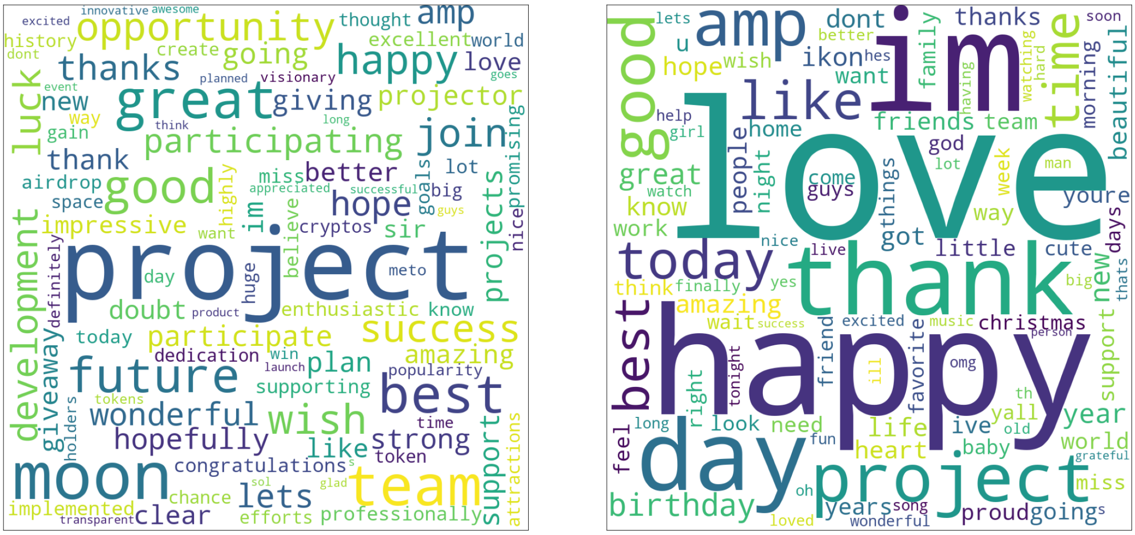
(b) Intuitive/Sensory

(c) Thinking/Feeling

(d) Judging/Perceiving

Figure 3.13: Variable Importance Plots (VIPs) based on an upsampled logistic regression model for each dichotomy – we include only EMOJI counts in the model. We determine the variable importance using the t-statistic for the parameters associated with each feature. Variables are sorted by the absolute value of the variable importance (from top to bottom). We colour the bar based on the features preference for each dichotomous class.

The rocket ship emoji is in the top 12 most important predictors for each of the four models. We observe that an increase in this emoji's usage makes an account more likely to be introverted, intuitive, feelings-orientated and perceiving. Recently, the rocket ship



(a) Rocket Ship Emoji

(b) Red Heart Emoji

Figure 3.14: Word clouds based on the most frequently occurring words in tweets from our dataset. Larger words appeared more frequently in the tweets - we present results for the rocket ship emoji (left) and the red heart emoji (right). Note that we remove stopwords as they do not provide much context for the tweets.

emoji has been taken over by finance enthusiasts who use the emoji to denote a fast increase in a particular stock or crypto-currency. Hence, it is likely that we are observing crypto enthusiasts to be more introverted, intuitive, feelings-orientated and perceiving. However, obviously this emoji has other meanings like its use to denote an actual rocket ship, so we performed further investigation to discover the general discussion of tweets using the emoji. We did this by creating word clouds of the words that appeared in tweets containing the emoji. These word clouds were based on the frequency of the words, so larger words appeared more frequently in the tweets. Note that we removed stopwords as they do not provide much context for the tweets. We display the results for the rocket ship emoji and the red heart emoji in Figures 3.14a and 3.14b, respectively. From these results, we observe the rocket ship emoji generally appears in tweets discussing ‘projects’, ‘great opportunities’, ‘developments’ and ‘cryptos’, validating this emoji is often used when discussing crypto. However, this emoji also sometimes appears in tweets discussing the ‘moon’ and ‘space’. Moreover, we see that the red heart emoji mainly appears in tweets discussing ‘love’ and ‘happiness’, validating this emoji is a good proxy of describing these attributes about a person.

Figure 3.13a shows a number of the emojis making an account more introverted are sad/upset emojis, whereas there are no sad/upset emojis which make an account more extroverted. This further confirms our discussion of Figure 3.12a which suggested that extroverts are more likely to display positive emotion online. Similar results have been discovered by other authors analysing emoji usage and personality types. For instance, Li *et al.* [90] found that extroversion was negatively correlated with the use of sad/upset emojis. Considering the Intuitive/Sensing results in Figure 3.13b, we observe that sensors tend to use more emojis with exaggerated facial expressions and rich emotions. Perhaps sensors use these types of facial expressions because "they are more concerned with what is actual, present, current, and real", as suggested by The Myers-Briggs Foundation [146]. Figure 3.13c shows that the most important emoji for the thinking trait is the thinking emoji, further confirming our results. We also observe a lot of the emojis making an account more feelings-orientated involve hearts. Perhaps this is because accounts exhibiting the feelings trait seek harmony and displaying these types of emojis are their attempt to create conformity online. Considering the Judging/Perceiving results in Figure 3.13d, two of the three most significant emojis for the Judging category are exaggerated laughing emojis.

So far we have used this section to determine the importance of individual predictors in our models – excluding the BERT features. Next we wanted to consider the importance of different feature groups (including the BERT features) and discuss whether different groups of features are more informative in our models. Again, we do this by fitting a logistic regression model to all of the features and performing the stepwise feature selection on each of the models. We use the same thresholds in our models of 0.05 and 0.1 for accepting and removing features, respectively. Recall for each model there are a total of 866 features to choose from: 11 SM features, 74 LIWC features, 768 BERT features, seven Botometer features and six VADER features. We then use the number of remaining features in each feature group after stepwise selection to measure the importance of the different feature groups. Table 3.9 displays the number of predictors in each feature group that remain in each model after the stepwise feature selection algorithm. Note that we also include the proportion of features in each group that are retained after stepwise selection. This proportion can be considered a measure of the importance of each feature group which is not biased by the number of features in each group – it can also be interpreted as the payoff per feature or the amount of information about Myers-Briggs personality types per feature. Furthermore, we introduce a more robust statistical framework to determine whether different groups of features are actually more informative of our data. We do this by performing a Chi-Squared Test on the number of features retained and excluded from each model – the Chi-Squared Test Statistic is defined in Section 2.4.2. We use the Chi-Squared Test to test the null hypothesis that each feature group is equally as informative (per feature) and include the p-values from the Chi-Square Test in the sub-table captions displayed in Table 3.9.

Table 3.9 shows that the number of features selected in each model is dependent on the type of model. For instance, 243 features are selected in the Intuitive/Sensory model, whereas only 124 features are selected in the Judging/Perceiving model. Interestingly,

Feature Type	Features Selected	Proportion Retained
SM	4	36.4%
LIWC	15	20.3%
BERT	176	22.9%
Botometer	1	14.3%
VADER	2	33.3%
Total	198	22.9%

(a) Extrovert/Introvert ($p = 0.720$)

Feature Type	Features Selected	Proportion Retained
SM	7	63.6%
LIWC	18	24.3%
BERT	217	28.3%
Botometer	0	0.00%
VADER	1	16.7%
Total	243	28.1%

(b) Intuitive/Sensory ($p = 0.032$)

Feature Type	Features Selected	Proportion Retained
SM	5	45.5%
LIWC	11	14.9%
BERT	124	16.1%
Botometer	1	14.3%
VADER	3	50.0%
Total	144	16.6%

(c) Thinking/Feeling ($p = 0.019$)

Feature Type	Features Selected	Proportion Retained
SM	4	36.4%
LIWC	8	10.8%
BERT	112	14.6%
Botometer	0	0.00%
VADER	0	0.00%
Total	124	14.3%

(d) Judging/Perceiving ($p = 0.120$)

Table 3.9: The number of features and the proportion of features retained in each feature group after performing the stepwise feature selection on all types of features. For each model, we perform the stepwise feature selection based on the p-values of the features; using an acceptance threshold of 0.05 and a removal threshold of 0.1. Moreover, we display the total number of features selection and the total number of features retained in each model. For each model, we perform a Chi-Squared Test to test the null hypothesis that each feature group is equally as informative (per feature) and display the p-value in the table captions. Note that there are a total of 866 features: 11 SM features, 74 LIWC features, 768 BERT features, seven Botometer features and six VADER features.

the Intuitive/Sensory model is also the most accurate and the Judging/Perceiving model is the least accurate – there is a positive relationship between accuracy and the number of features retained in the model. This is consistent with the remark that more features are retained in a model when they are more informative about the data. Moreover, the SM features are on average the most retained features across the models, as these features seem to have the best payoff across the four models. Conversely, the Botometer features have the worst payoff across the four models as they have the smallest proportion retained, on average. This is likely due to the Botometer features being highly correlated (see Figure 3.5), as they are all inferring similar information about the user. The most interesting comparison is between the LIWC and BERT features, as both of these features aim to describe linguistic properties about an account's text usage. In each model, the BERT features are more highly retained. However, only the results from the Intuitive/Sensory model and the Thinking/Feeling model are significant at the 5% level. We therefore reject the null hypothesis that each feature group is equally as informative (per feature) for the Intuitive/Sensory and Thinking/Feeling models – indicating that some feature groups are more informative than others in these models. However, the Chi-Squared Test does not alone tell us what feature groups perform significantly better. So we perform individual confidence intervals for the binomial proportions of accepting/rejecting features in each group. In each case, we use the Wilson Score interval [127], a corrected version of confidence intervals for binomial proportions with zero successes, seen in Table 3.9. The confidence intervals for each feature group and each model are displayed in Figure 3.15.

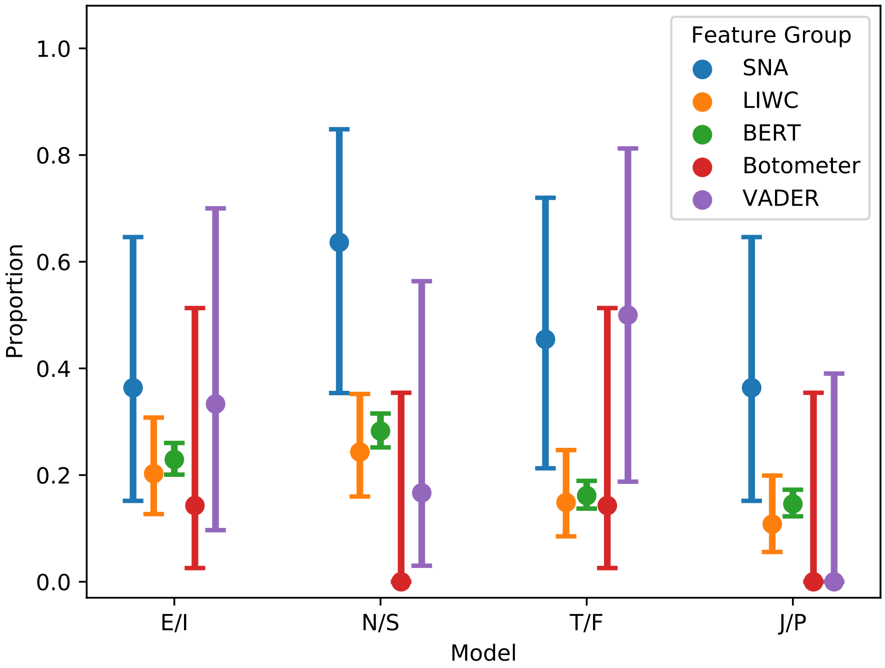


Figure 3.15: 95% Wilson Score Binomial Confidence Intervals for the proportion of retained features in each feature group. We display the confidence intervals for each model and use the Wilson Score version to correct for having zero successes in some cases. The confidence intervals are based on the results in Table 3.9 and we observe they align with the p-values observed in this table.

Figure 3.15 shows that the Intuitive/Sensory and the Thinking/Feeling models are the only two models with non-overlapping confidence intervals. This aligns with the p-values from the Chi-Squared Test (see Table 3.9). For the Intuitive/Sensory model, the 95% confidence interval for the SM features lies completely above the 95% confidence intervals for the LIWC and BERT features. This indicates that the SM features are more informative (per feature) than the LIWC and BERT features at the 5% level. This highlights that attributes about a user's account are sometimes more important than the language they use when modelling personality. Hence, it is important that other researchers use these account level attributes when modelling the personality types of social media users. Adding to this, we observe that the 95% for the SM features and VADER features lies completely above the 95% confidence interval for the BERT features in the Thinking/Feeling model. This further supports our previous claims that the account level features are often more important (per feature) than the textual features for some of the models. Note that the textual features are all fairly correlated with each other, as observed in Figure 3.4. If so, then many of the textual features won't be added to the model because they provide similar information about an account. However, since we observe these effects only for two out of four models, it is likely that these textual features are simply less important for determining the Intuitive/Sensory and Thinking/Feeling traits.

Summary

The main contributions of this chapter included the labelled dataset of personality types we obtained and the tools/techniques which we introduced to model the personality types of these users. To our knowledge, this dataset is the largest available dataset of labelled Myers-Briggs Personality Types on Twitter. The only other labelled Twitter dataset (to our knowledge) was created by Plank and Hovy [117] and contained only 1,500 labelled accounts – making our dataset considerably larger. Moreover, the data collection techniques we used to collect this data are also novel, as they avoid the long, cumbersome questionnaires which other researchers have used. Moreover, we develop a statistical framework which combines NLP tools and mathematical models to model/predict the personality type of users online. While we have applied this framework to personality types, it can be more broadly utilised to model any labelled characteristics of online accounts – political opinions, psychological properties or even someone's propensity to adopt an opinion or viewpoint. As part of this framework, we analyse and compare a number of different machine learning models. Since personality types in our dataset are unbalanced, we compare different weighting/sampling techniques to deal with issues arising from class imbalance. We discover that SMOTE performs poorly when combined with naive Bayes and demonstrate the reasons for this by visualising the effects of both tools on a lower dimensional example. Finally, we compare the importance of different features in our models. We do this by comparing the effects of different features on someone's personality type at the individual and group level.

While this chapter provides a thorough analysis of our dataset as well as different personality models, there is certainly a need for future work in this area. Since we use a large number of features on a fairly large dataset, a deep learning model is certainly appropriate for this type of problem. Hence, it would be desirable to test the performance of our features on this dataset by utilising a suite of deep learning models. These may include models such as: Recurrent Neural Networks, Perceptron, Long Short Term Memory (LSTM) and a number of other more advanced black-box type machine learning models. Obviously these models would not give the same interpretability as the models we have used in our analysis, so they would be primarily used for their predictive capability. It would also be interesting to consider different methods for collecting data. One limitation of our dataset is that we only have access to the classification of the four personality dimensions, when in reality these dimensions are represented on a numerical scale. For instance, two users may be extroverted but one user may be considerably more extroverted than the other. While performing questionnaires are long and expensive, it would enable us to obtain these personality dimensions on a numerical scale, and it would reduce the mislabelling rate of the accounts. We would expect this to have a significant improvement on the performance of our models. Another obvious extension of our work is to use the OCEAN personality model instead of the Myers-Briggs model. By utilising questionnaires to obtain our data, we would have the luxury to choose which personality model to use, and so it would be possible to consider using the OCEAN model. We could then consider obtaining both personality types for each user and perform a comparison between the two personality models. This would enable us to test the reliability and accuracy of both personality models, something which has not been done by any other researchers (to our knowledge).

This chapter explains how people's online digital footprint can be utilised to understand personal attributes about them. However, we have not discovered how this type of information can be (mis)used at large scales. One current application of personality profiling is to use it in information warfare campaigns and target people with campaigns based on their personality types. This potentially gives microtargetting groups the ability to manipulate public opinion at large scales when they have access to this information. As a result, in the next section we will develop a statistical framework for measuring how public opinion can be impacted at large scales. We will do this in the context of the 2022 Russian invasion of Ukraine, emphasising the role that social media plays in modern-day conflicts.

Chapter 4

The interaction of bots and humans in discussion of the Russia/Ukraine war on Twitter

The research in this chapter was published as part of the following paper;

Smart, B., Watt, J., Benedetti, S., Mitchell, L., Roughan, M. [2022], #IStandWithPutin Versus #IStandWithUkraine: The Interaction of Bots and Humans in Discussion of the Russia/Ukraine War, *in* F. Hopfgartner, K. Jaidka, P. Mayr, J. Jose and J. Breitsohl, eds, 'Social Informatics', Vol. 13618 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, pp. 34-53. `https://doi.org/10.1007/978-3-031-19097-1_3`.

This chapter is an expanded version of Section 4 and 6 of the above paper by Smart *et al.* [139]. In this chapter, we explore the extent to which Twitter can be weaponised to influence public opinion at large scales. We develop a statistical framework for measuring the effects of bot-like accounts on the online discussion in the context of the Russia/Ukraine war. We do this by studying the #IStandWithPutin campaign versus the #IStandWithUkraine campaign, emphasising the role which social media plays in modern-day conflicts. While we consider Twitter in this analysis, the statistical framework we develop can be more generally applied to any social media application and any campaign we wish to measure the effects of.

Both sides in the Russia/Ukraine conflict use the online information environment to influence geopolitical dynamics and sway public opinion. Russian social media pushes narratives around their motivation, and Ukrainian social media aims to foster and maintain external support from Western countries, as well as promote their military efforts while undermining the perception of the Russian military. Examples of these narratives include allegations: that Ukraine was developing biological weapons [161], that President

Volodymyr Zelenskyy had surrendered [28, 85], and that there is a sustained campaign showing the apparent success of ‘The Ghost of Kiev’ [89]. Some of the information being pushed is genuine, and some is malicious. It is not easy to discriminate which is which.

As a result, we learn how bots are influencing the online conversation by measuring what communities are talking about online, and how this discussion evolves. As part of this, we collect a dataset of approximately 5.2 million tweets, retweets, quote tweets and replies over the first two weeks of the war that contain hashtags in support of Russia/Putin and hashtags in support of Ukraine/Zelenskyy. We use Linguistic Inquiry and Word Count (LIWC) to determine the linguistic features of the Twitter content and calculate the Botometer scores on a random sample of accounts in our dataset – details on LIWC and Botometer are provided in Sections 2.3.1 and 2.3.4, respectively. Moreover, we employ time-series analysis techniques on this data to understand how bot-like activity impacts the conversations surrounding the war. We create an approach that is robust, transferable and able to be quickly applied to large volumes of data. In essence we seek to determine whether the malicious influence campaigns work as intended.

4.1 Data

In this section, we discuss how we collected the data utilised in our research which underpins how bots and humans interacted during the beginning phases of the Russia/Ukraine War. We then provide the necessary steps used to preprocess the data - this includes the steps we used to remove accounts and the methods we used to calculate various attributes about accounts in our dataset. Finally, we perform an Exploratory Data Analysis (EDA) on the dataset to describe important features which arise in our data.

4.1.1 Data Collection

The Twitter API (V2) was utilised to obtain all Tweets, Retweets, Quotes and Replies that contain case-insensitive versions of the hashtags provided in Table 4.1. This Twitter content was obtained from February 23rd 2022 00:00:00 UTC until March 8th 2022 23:59:59 UTC and represents the fortnight (2 weeks) post Russia’s invasion of Ukraine (predating the invasion by one day). Note that we query the hashtags with and without the ‘I’ at the beginning, resulting in a total of 12 hashtags that are queried.

We chose these hashtags because they were found to be the most trending hashtags on Twitter which could be clearly associated with one side of the conflict. The resulting dataset contained 5,203,746 samples from Twitter and is comprised of 503,492 Tweets, 4,102,030 Retweets, 289,899 Quotes and 308,325 Replies. For each quote, we only obtain the text from the quote and not any text from the tweet being quoted. Whereas for the retweets, we obtain the text for the entire tweet being retweeted. We do this because a retweet is nearly always shared in agreement with the text in the original tweet, whereas

Pro Ukraine	Pro Russia
#(I)StandWithUkraine	#(I)StandWithRussia
#(I)StandWithZelenskyy	#(I)StandWithPutin
#(I)SupportUkraine	#(I)SupportRussia

Table 4.1: Hashtags which are queried to obtain our dataset (we refer to these as query hashtags). We use these hashtags to obtain all Tweets, Retweets, Quotes and Replies over the fortnight post Russia's invasion of Ukraine, predating the invasion by one day.

a quote could either be used in agreement or disagreement of the tweet being quoted. Consequently, a quote would only appear in our dataset if one of the query hashtags is contained in the quote itself.

4.1.2 Preprocessing

In this section, we discuss the preprocessing steps performed on the data in order to prepare it for analysis. This includes a removal of unwanted Twitter content, calculating LIWC scores, calculating Botometer scores, classifying accounts into bot types and calculating a national 'lean' for the accounts. We define the national 'lean' as an accounts preference for being in support of either Russia/Putin or Ukraine/Zelenskyy.

Our Twitter content removal process only involved one step, which was to simply remove all Twitter content which isn't recognised as English-spoken. We do this because the LIWC scores can only be calculated on English-spoken language. Moreover, we detect these languages using Twitter's language detector; which calculates the language of a Tweet, Retweet, Quote or Reply. Twitter's language detector can detect over 100 different languages and we find that our dataset contains 65 of these detectable languages. The most commonly occurring language in our dataset in English, with 3,524,776 posts (67.7%) being written in English – forming the resulting dataset we used in the remainder of our analysis. Moreover, 10.8% of the posts came from an unknown language and, Ukrainian and Russian Language each contribute to approximately 1% of the content. The top ten most frequently occurring languages in our dataset are included in Table 4.2.

In Table 4.2, we observe there to be a significant amount of English-spoken content compared to any other language. This is significantly more English-spoken content than what is generally observed across the Twittersphere – for instance, a study by Hong *et al.* [70] found only 51.1% of the content on Twitter is written in English. We suspect this is because the hashtags we queried were in English and so our sample is biased to containing more English spoken content. However, we are not concerned with this bias in our data because we only study English content in all of our remaining analysis. Moreover, we also observe there to be significantly more Ukrainian and Russian spoken language in our dataset, compared to the study by Hong *et al.* [70]. This is indicative that Russian and

Language	Twitter Content	Proportion
English	3,524,776	67.7%
Unknown	563,617	10.8%
Japanese	255,937	4.91%
Thai	157,202	3.02%
German	125,046	2.40%
Spanish	84,135	1.62%
French	81,139	1.56%
Polish	72,207	1.39%
Ukrainian	52,419	1.00%
Russian	51,009	0.980%

Table 4.2: The top ten most frequently occurring languages of content in our dataset. The total dataset is comprised of 5,203,746 posts and was obtained using the query hashtags given in Table 4.1.

Ukrainian people are much more likely to contribute to the online dialogue surrounding the Russia/Ukraine War – compared to other topics – highlighting this topics importance to both groups of people.

After removing all non-English spoken content, we calculated the LIWC scores on each post. Information on LIWC can be found in Section 2.3.1 and we have previously included all of the LIWC features in Table 3.4. Recall that LIWC uses a one-to-many mapping for words to LIWC categories, so words will often appear in several categories. As a result, we calculate the LIWC features for each post by finding the proportion of words in each of the LIWC categories. We can interpret these scores as the proportion of words from a post that convey the particular LIWC emotion/category.

Next we calculated the Botometer scores for a sample of these accounts. We calculated all of the Botometer scores for these accounts – these Botometer scores can be found in Table 3.4. However, we only utilise the language-dependent Botometer scores in any of our analysis; we are able to do this because we are only considering English based content. Another reason for doing this is because the language-dependent Botometer scores have been shown to be more accurate [126]. The Botometer rate limits allowed us to randomly sample 483,100 (26.5%) unique English accounts in our dataset which posted at least one English Tweet. We note that this random sample leads to an approximately uniform frequency of Tweets from accounts with Botometer labels across the time frame we considered. Due to rate limit constraints, the Botometer scores were calculated post-collection, so a small number of accounts may have been removed or scores may be calculated using activity after our collection period. We also acknowledge that Twitter's takedown of Russian accounts on the 3rd of March may lead to less bots in our dataset.

However, analysis showed that the content spread by these accounts persisted despite the takedown[1].

The query hashtags from each tweet were extracted and the total number of pro-Ukrainian (ending in Ukraine or Zelenskyy) and pro-Russian (ending in Russia or Putin) hashtags were counted and used to establish the national 'lean' of a tweet. If the number of pro-Ukrainian query hashtags exceeded that of the pro-Russian hashtags, the tweet was labelled as 'ProUkraine', and labeled as 'ProRussia' conversely. If the counts were balanced, the tweet was labelled 'Balanced'. Where applicable, the lean of an account was taken to be the most commonly occurring national lean across all tweets from that account. We show the percentage of account leans in Table 4.3, where we found that 90.16% of accounts fell into the 'ProUkraine' category, while 6.80% fell into the 'ProRussia' category. The balanced category contained 3.04% of accounts, showing that accounts exhibiting mixed behaviour are present in the dataset.

National Lean	Percentage
ProUkraine	90.1567
ProRussia	6.7988
Balanced	3.0444

Table 4.3: Proportion of Accounts labelled with each national lean label. This shows that most accounts fell into the 'ProUkraine' category and a surprising number had most of their tweets containing a balanced number of 'ProUkraine' and 'ProRussia' hashtags.

Note that we explored other methods for categorising accounts, e.g., labelling accounts as 'ProUkraine' or 'ProRussia' if they use only those types of hashtag. However, as we were primarily concerned with aggregated activity, we elected to prioritise labelling each account by their 'usual' behaviour.

The resulting dataset[2] of Twitter users who participated in discussions around the Russian Invasion of Ukraine was published to Figshare [158].

4.1.3 Exploratory Data Analysis

We perform an Exploratory Data Analysis (EDA) of our dataset in order to explore the different features and provide an overview of our data. Firstly, we wanted to explore the activity of bots over our considered timeframe as well as how this aligned with the usage of different hashtags and Twitter content types. We present the results in Figure 4.1, separated into three different subplots. The top subplot describes the average hourly

[1] `https://twitter.com/timothyjgraham/status/1500101414072520704`

[2] Dataset available at `https://figshare.com/articles/dataset/Tweet_IDs_Botometer_results/20486910`.

proportion of bot types in our dataset over the considered time frame. Note that the 'Overall' bot type refers to the Complete Automation Probabilities (CAPs). The middle subplot describes the hourly frequency of hashtags in Table 4.1 over the considered timeframe. And, the bottom subplot described the hourly frequency of Twitter content over the considered timeframe. In these figures, we also include some of the main events over the first fortnight post Russia's invasion of Ukraine and observe how these events shaped discussions on Twitter. These events include: when the conflict begins (24th February 2022), when the fighting in Mariupol begins (26th February 2022), when Russia captures Kherson (2nd March, 2022), when Russia captures the Zaporizhzhia nuclear power plant (4th March 2022) and when Ukrainian authorities first attempt to evacuate Mariupol (8th March 2022).

We can observe the clear presence of a daily cycle in the average hourly bot probabilities in Figure 4.1. However, it is important to recognise that the daily cycle with the Astroturf accounts seems to be opposite to the daily cycle with all other types of bots. This observation is further validated in Figure 4.2, where we average across Botometer probabilities based on the hour of day (UTC time) and centre around the mean. We believe Astroturfing accounts are active at opposite times due to two potential reasons; either the Astroturfing accounts are from a different timezone to a majority of the accounts or some property of Botometer is using the timezone to determine whether an account is Astroturfing. It is also important to observe the spike in bots on the 2nd and 4th of March. This first spike aligns with when Russia captured the first Ukrainian city, but also when the #(I)StandWithPutin and #(I)StandWithRussia hashtags were trending. Hence, further justifying our belief that certain bot armies were responsible for making these hashtags trending. The second spike in bot activity on the 4th of March is more difficult to justify. The 4th of March represents when Russia captured the nuclear power plant but also when a handful of pro Russian accounts were removed for violating Twitter's policies. Perhaps this spike in bot activity may be due to the presence of pro-Ukrainian bots which are advocating against the pro-Russian accounts. Nonetheless, there is an obvious presence of bots over the duration of the first fortnight post Russia's invasion of Ukraine.

We also observe a clear presence of a daily cycle in the hashtag frequencies. Further to this, we observe an initial spike in the #(I)StandWithUkraine tweets, with this hashtag also being the most dominant over the considered timeframe. This signifies the world's support of Ukraine when the conflict began and throughout the initial phases of the war. Interestingly, we observe a spike in the #(I)StandWithPutin and #(I)SupportPutin hashtags on the 2nd and 3rd of March, just after Russia captures its first Ukrainian city. We believe these spikes in support of Putin are predominately due to the presence of bots, an observation also made by researcher Timothy Graham in an article posted by Purtill [124]. However, Twitter took action on these accounts on the 4th of March by removing over 100 users which pushed the #(I)StandWithPutin campaign to become a trending

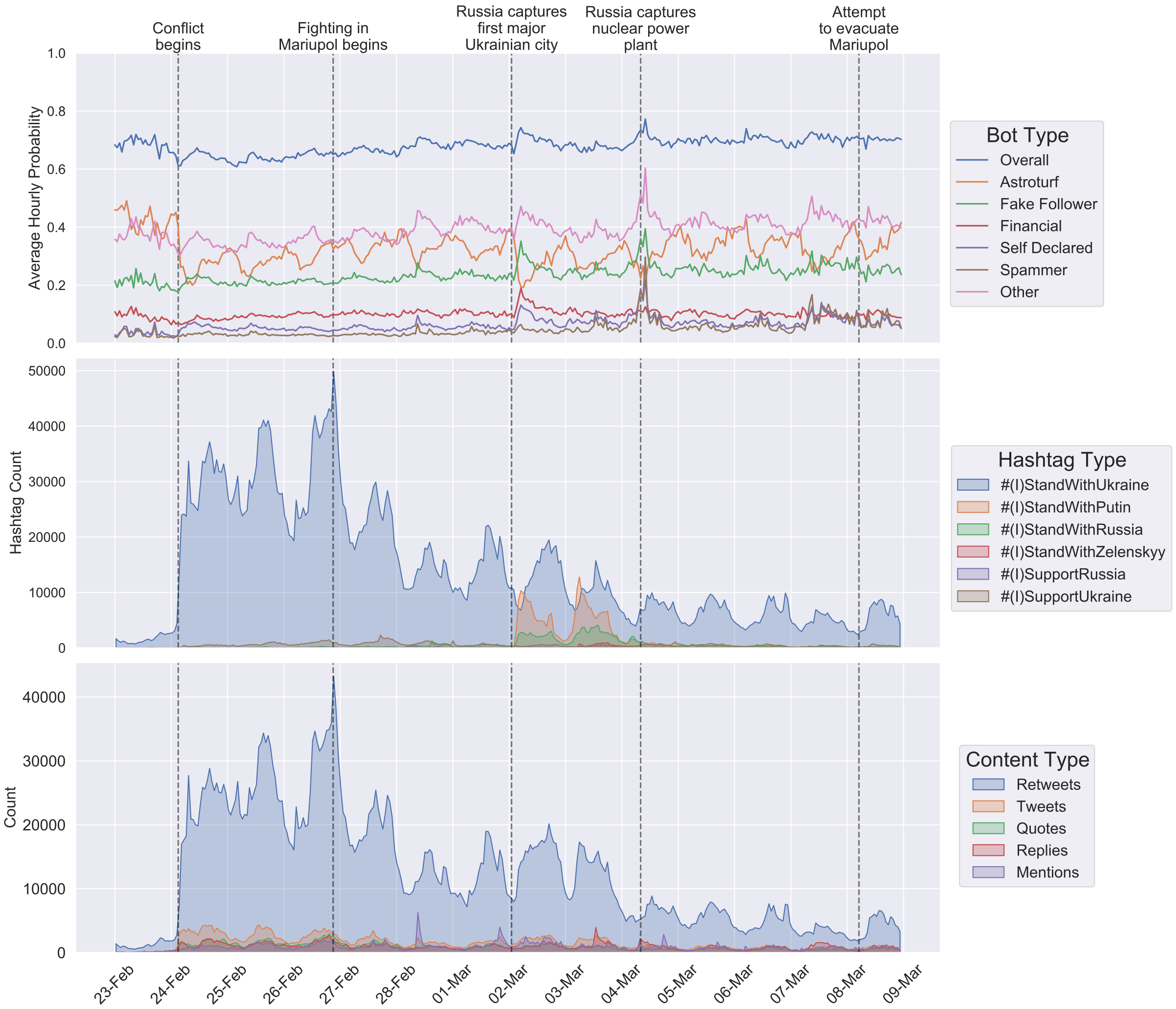


Figure 4.1: Average hourly probabilities of bots tweeting query hashtags (top). Hourly frequency of the query hashtags (middle). Hourly frequency of content (bottom). The time period we consider is the first fortnight after Russia's invasion of Ukraine. Both plots also include five significant events over this time period. Note that the query hashtags can be found in Section 4.1.1. We can observe a significant spike in the bot activity of several bot types on the 2nd and 4th of March. The spike in bot activity on the 2nd of March aligns with Russia's capture of Kherson, and also aligns with a significant increase in pro-Russia hashtags. This spike in activity was due to an increase in activity of pro-Russian bots – likely used by Russian authorities. The spike in bot activity on the 4th of March aligns with when the use of pro-Russia hashtags diminished, but also when Russia captured the Zaporizhzhia nuclear power plant. This spike was due to an increase in activity of pro-Russian bots (before being removed) and an increase in activity of pro-Ukrainian bots – likely by pro-Ukrainian authorities in response to Russian bots.

topic and violated its “platform manipulation and spam policy” [34]. As a result, we can observe this campaign diminishing from the 4th of March.

The daily cycle can also be observed in the content frequencies, where we see that retweets are consistently the most used content type over the timeframe we considered. Furthermore, we observe spikes in all forms of content when the fighting in Mariupol begins, marking the most content we observe in a single day over the fortnight we considered. We also see noticeable spikes in the mentions on 28th of Feb, 1st of March and 4th of March – this could be due to there being a lot of online conflict on these days.

Next we considered the activity of different bot types based on the time of day, something we alluded to earlier in this section. In Figure 4.2, we present the average Botometer probabilities, averaged based on the hour of day (UTC time) and centered around the mean. The coloured range for each of the bot types represents the 95% confidence interval for the data at each coordinate.

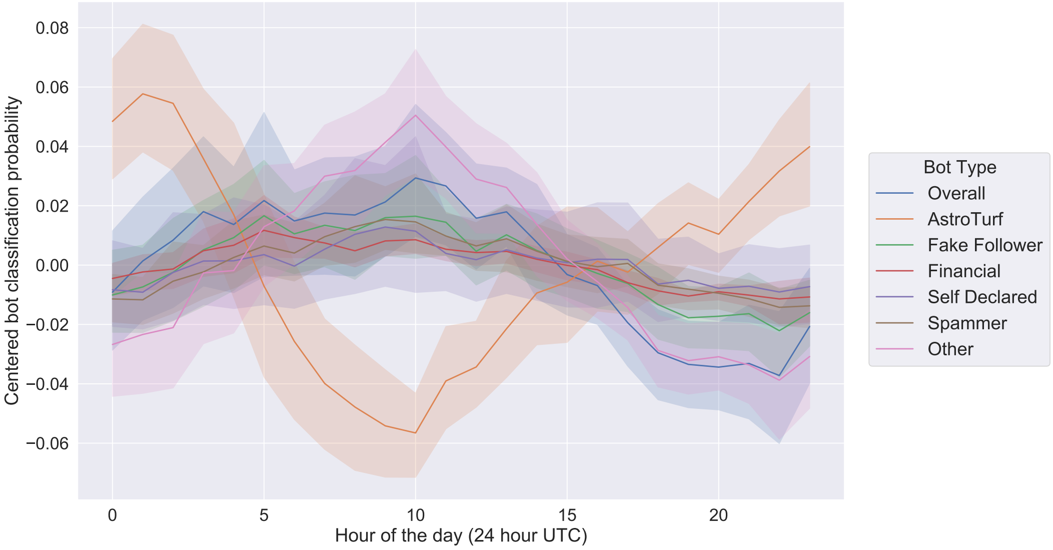


Figure 4.2: Average hourly Botometer results showing the daily cycle. The time series observed in Figure 4.1 (top) is averaged based on the hour of the day (UTC time). Note that the coloured range for each of the bot types represents the 95% confidence interval for the data at each coordinate.

In Figure 4.2, we can further observe the effects of time of day to be most pronounced for the activity of Astroturfing and Other bots. Whereas, the activity of Fake Follower, Financial, Self Declared and Spammer bots seem to be less impacted by the time of day. One reason for this may be because Astroturfing and Other bots are pushing campaigns specific to certain countries and hence it it most practical to share the content at a certain time. In regard to the Other bots, we can observe the spike to occur at 10:00 UTC time which corresponds to 1:00pm Ukrainian time. Note that this is also when the 'Overall' bot probability is highest, indicating this is the time when the bots are generally most active. Further to this, Matthews [93] suggested that Noon to 1:00pm is the most popular time to tweet in any timezone. Hence, the bots are likely to be increasing their engagement in Ukraine by being most active around this time.

Next, we consider how the probability of different bot types varied based on their national lean. We did this by producing box plots of the bot probabilities based on an accounts national lean and bot type – the results are displayed in Figure 4.3.

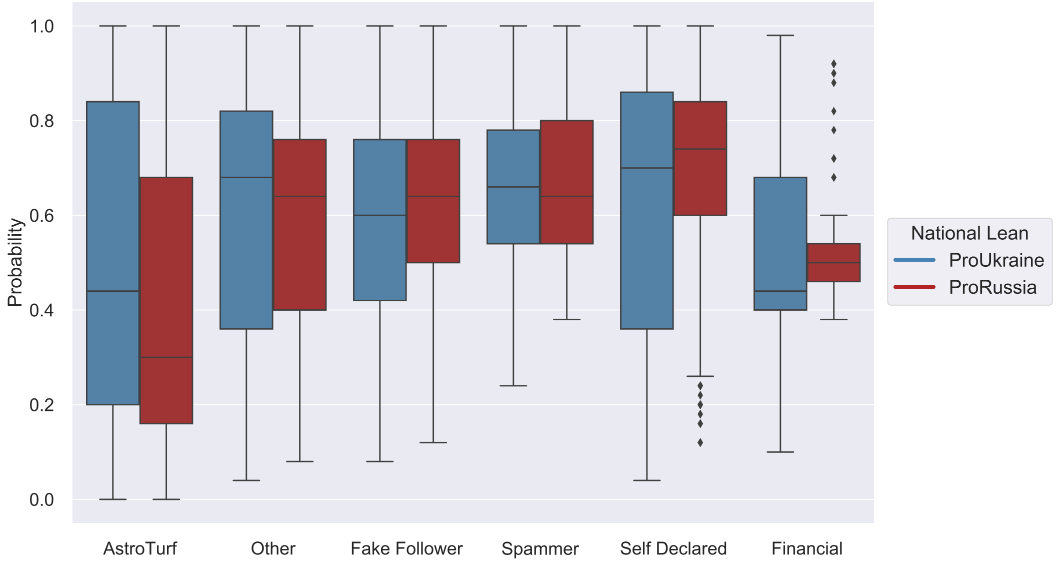


Figure 4.3: Probabilities of bot types based on national lean and bot type. Establishing the national lean is described in Section 4.1.2.

Figure 4.3 shows that the most commonly used bot type for both campaigns is self declared bots. This suggests authorities have identified these bots to be most useful in a information warfare campaigns. Furthermore, we can observe a fairly consistent spread of bot types for both campaigns. This indicates that bots are consistently spread across both campaigns, and there is not one campaign utilising significantly more bots than the other. Pro-Russian accounts have a mean Complete Automation Probability (CAP) score of 0.42, while pro-Ukrainian accounts have a mean score of 0.43, with medians 0.36 and 0.34 respectively. This further validates that both campaigns are utilising automated accounts on Twitter approximately equally. However, the median probability of an account being an Astroturf bot is slightly higher for pro-Ukrainian accounts than pro-Russian accounts. Additionally, the median probability of an account being a self-declared bot is slightly higher for pro-Russian accounts compared to pro-Ukrainian accounts. This highlights that pro-Ukrainian authorities may be utilising more Astroturfing accounts in their information warfare, whereas pro-Russian authorities may be utilising more self-declared bots.

Finally, we consider whether there is noticeable differences in the usage of bots based on the national lean of accounts. We do this by producing distributions of the 'Overall' bot probabilities (CAP) based on the national lean of accounts. Note that the distributions are normalised separately because of the imbalance in pro-Ukraine and pro-Russia account groups. The results are displayed in Figure 4.4.

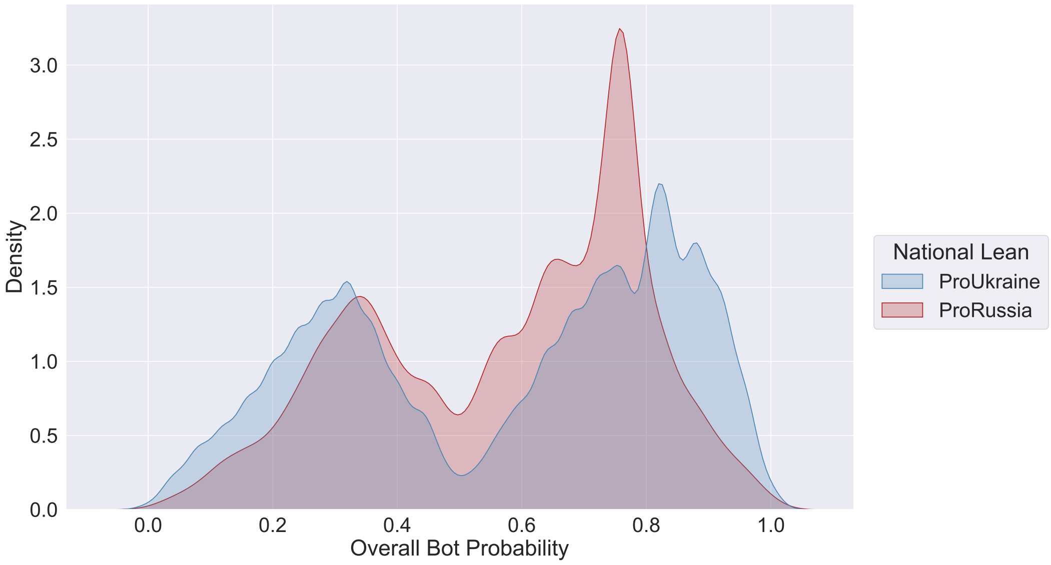


Figure 4.4: The distribution of Overall bot probabilities based on the National Lean of accounts. Both distributions appear bimodal, signifying there are very few accounts exhibiting both bot-like and human-like behaviour.

The distributions in Figure 4.4 are both clearly bimodal, something which was also observed in other studies of Botometer [163]. The bimodal shape of both distributions clearly show a group of users which are highly likely humans and a group of users which are very likely to be bots. Furthermore, we observe the overall probabilities of pro-Ukrainian accounts to have heavier tails compared the overall probabilities of pro-Russian accounts. This indicates that Botometer is more confident in predicting both human-like and bot-like behaviour in the pro-Ukrainian accounts compared to the pro-Russian accounts. However, this may also indicate that the pro-Russian bots are harder to detect than the pro-Ukrainian bots – perhaps because the authorities creating these bots are better at hiding their bot-like behaviour. It is worth noting the large peak in pro-Russian accounts around 0.78. This peak indicates there are a lot of pro-Russian accounts which are exhibiting similar amounts of bot-like behaviour. Moreover, it is also worth noting the peak in pro-Ukrainian accounts around 0.81. While this peak is not as prominent as the peak in pro-Russian accounts, it also demonstrates that there are groups of pro-Ukrainian accounts which are exhibiting similar amounts of bot-like behaviour. Note that we avoid commenting on the numbers of bot accounts and human accounts supporting both campaigns – this is because of the significant imbalance of accounts supporting each campaign. Because of this imbalance, the normalised distribution of accounts is a fairer comparison of bot-like behaviour exhibited by both sides of the conflict.

4.2 The Effects of Bots on the Discussion

In this section, we aim to discover how different types of bots drive feelings and discussions around the Russia/Ukraine conflict on social media. As part of this, we firstly consider how the discussion topics of bots may differ to the discussion topics of human accounts. We calculate the correlation between the bot probabilities of a tweets author and the LIWC category proportions of the tweet. The results are then displayed in a heatmap, which can be seen in Figure 4.5.

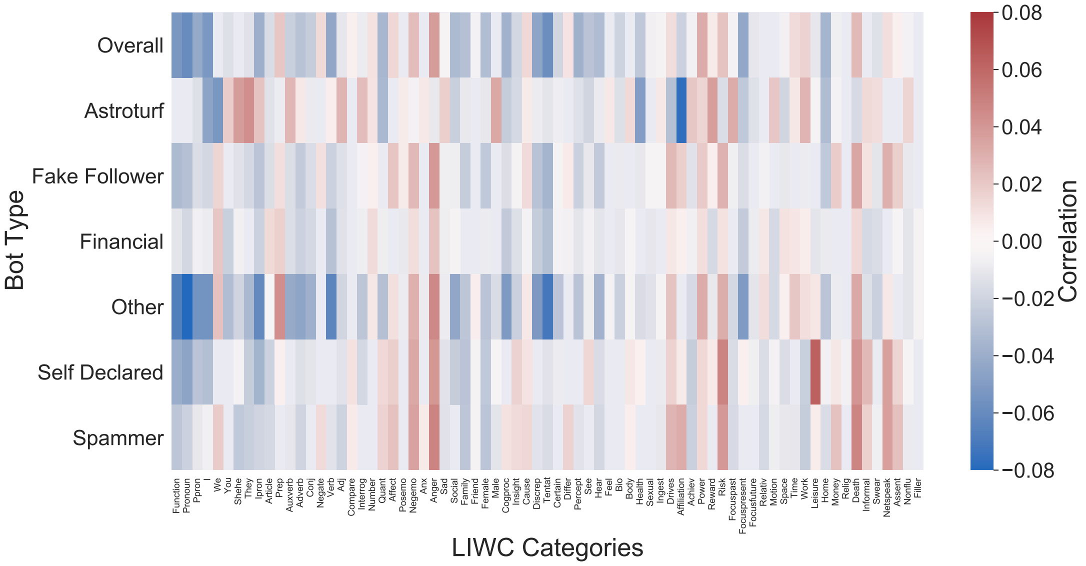


Figure 4.5: The correlation between different bot types and each of the LIWC categories. The panel of each square represents the correlation strength between the two variables. This figure can be used to understand whether we observe different conversations among the bots, compared to human accounts. We observe fairly low correlations in each case, with all being no greater than 0.1 and no less than −0.1.

From the results in Figure 4.5 convey fairly low correlations between each of the bot types and the LIWC categories. In particular, we observe the correlations to be no greater than 0.1 and no less than −0.1. This indicates very little changes in the linguistic content produced by the bot accounts and the human accounts. It is also important to note that the authors of Botometer do not publish their model architecture or report the features included in their model. Hence, there is a possibility that Botometer could be utilising the LIWC features (or linguistic features similar to LIWC) in their model – particularly the language-dependent Botometer model (which we used in our analysis). However, the low correlations in Figure 4.5 indicate that it is unlikely that linguistic features like LIWC are included in the architecture, and if they are then it appears they are not very useful for distinguishing bot accounts to human accounts.

Now that we have established there is very little difference in the linguistic content used by bots and humans, we next wanted to consider the effects of bot activity on the linguistic content of the online discourse on Twitter. Consequently, we produce hourly averages for the LIWC proportions and the Botometer probabilities for all content posted within each hour. This results in a set of time series, over 336 hours. We check for stationarity in the time-series data using an Augmented Dickey-Fuller Test [42] and discover that only the time-series for the 'other' bot probabilities does not pass the test of stationarity. We consequently enforce stationarity in the time series for these bot probabilities by removing the linear trend from this data.

We utilise the Granger Causality Test on these time series to determine whether the activity of certain bots Granger-cause more/less discussion around particular LIWC categories – Granger Causality was discussed in Section 2.4.2. We fit the Granger Causality Test over 12 time lags (hours) with the LIWC categories being the response variable. We use 12 time lags because we believe it is reasonable to assume a majority of the effects from bots will occur over this time frame. However, we will consider the validity of this assumption in more detail below (see Figure 4.8). We use the F-score from the Granger Causality Test as a measure of how 'influential' a type of bot is on discussions around each LIWC category. To get a sense of direction for these relationships, we use the sign of the largest β coefficient from Eq. 2.2 in Section 2.4.2. We multiply the sign of this coefficient by the F-score from the Granger Causality Test to obtain a measure of strength and direction, and refer to this as the 'Bot Effect Strength & Direction'. Moreover, we use the lag of the largest β coefficient from Eq. 2.2 in Section 2.4.2 to represent the most prolific lag in the relationship. We use the p-value from the Granger Causality Test to determine whether the effects are significant and perform a Bonferroni Adjustment to adjust for multiple hypothesis tests. The results are displayed in Figure 4.6, where we have only included the significant relationships. The number in the centre of each square represents the most prolific lag – we interpret this as the number of hours until the effects of the bot activity are most pronounced.

In Figure 4.6, we can observe the bots do have a significant impact on discussions of certain LIWC categories. To gain further context around what each of these LIWC categories represent, we generated word clouds of the words appearing in each LIWC category. Note that the size of the words represent their relative frequency in the data - the

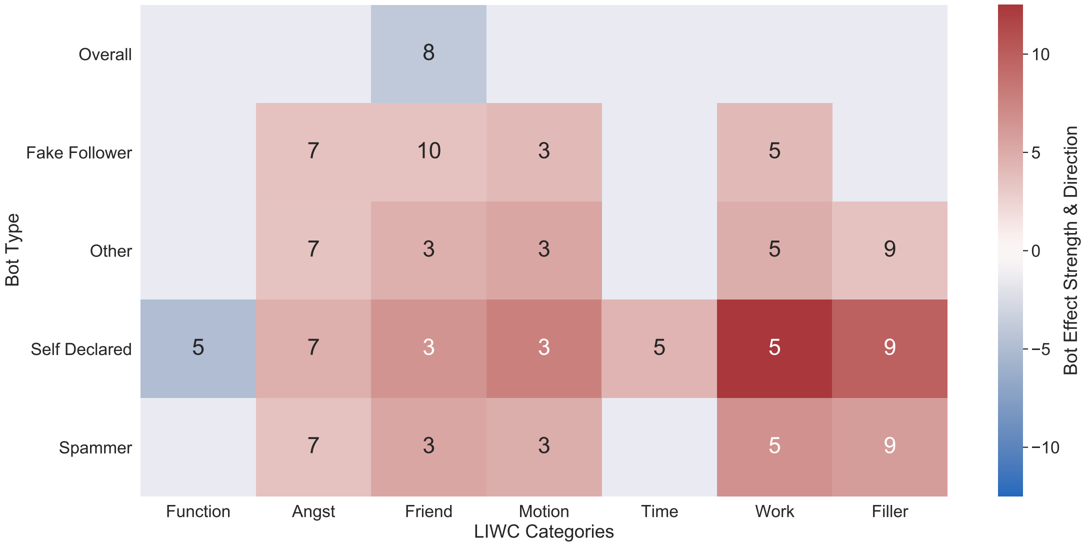
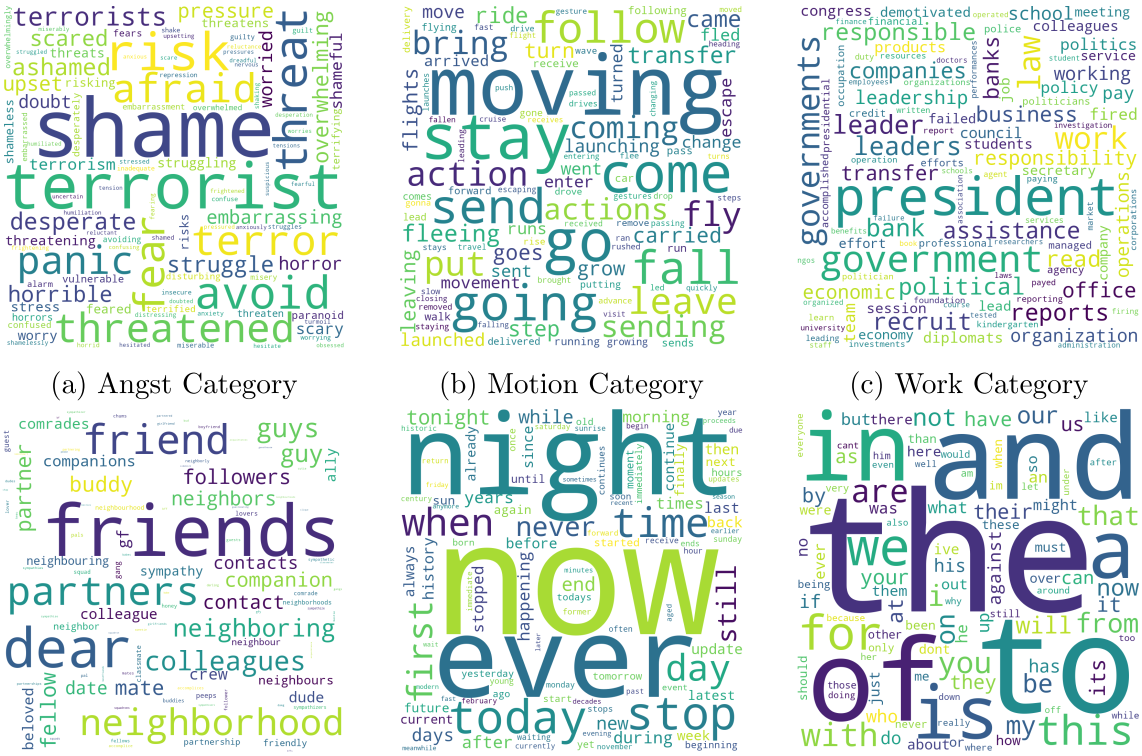


Figure 4.6: A series of pairwise Granger Causality Tests are performed to examine whether the activity of bot types is Granger-causing changes in discussions of the LIWC categories. The heat maps colour describes the bot effect strength and direction from the Granger Causality Test (over 12 hours/lags) between the time series of hourly bot proportions and the time series of hourly LIWC category proportions. The number in the centre describes the most prolific lag in the Granger Causality Test. We calculate the bot effect strength using the F-score from an F-test on the Granger Causality linear models. Moreover, we calculate the bot effect direction and most prolific lag using the sign and lag (respectively) of the largest β coefficient from Eq. 2.2 in Section 2.4.2. We perform a Bonferroni Adjustment on the p-values from the Granger Causality Tests and only show the Bot Types and LIWC Categories with a significant adjusted p-value (< 0.05).

larger the word the more frequently it occurs. These word clouds are provided in Figure 4.7, where sub-figure 4.7a, 4.7b, 4.7c, 4.7d, 4.7e and 4.7f represent the Angst, Motion, Work, Friend, Time and Function categories, respectively. Note that a full discussion of the words associated with various LIWC categories is provided in [114].

In Figure 4.6, the self declared bots seem to be having a great amount of influence on a number of discussions. In particular, we can observe the self declared bots increase discussions around angst, friends, motion, time, work and the usage of filler words but decrease the usage of function words. Moreover, it is apparent that the self declared bots are most strongly influencing discussion of the work category (with the prolific lag after 5 hours). In Figure 4.7c it appears most of the discussion around work is involved with governing bodies - with 'president' and 'governments' being the most commonly used words. While it is difficult to assert exactly why these bots are Granger-causing more discussions of work, we gain further understanding by also observing that self declared



(a) Angst Category

(b) Motion Category

(c) Work Category

(d) Friend Category

(e) Time Category

(f) Function Category

Figure 4.7: Word Clouds which demonstrate the frequency of words in particular LIWC categories. Larger words appear more frequently, relative to smaller words in the word cloud.

bots are Granger-causing more angst related discussion (with the most prolific lag after 7 hours). By combining these two observations, it is possible that self declared bots are driving more angst about governing bodies. From a pro-Russian perspective, this may be to cause more disruption in the West and from a pro-Ukrainian perspective, this may be to cause more disruption in Russia. Figure 4.3 shows there to be a fairly even probability of a pro-Russian account and pro-Ukrainian account being a self declared bot. Although the exact origin of the self declared accounts is unknown, it is worth noting that all the accounts we considered in this analysis are predominately English; so it is more likely that the intention of these accounts was to drive more disruption in English speaking countries.

Observe that Fake Follower, Spammer and Other bots also Granger-cause an increase in angst (all with the most prolific lag after 7 hours). From Figure 4.7a, we can observe that a majority of the angst-related words are related to fear and worry – suggesting that self declared, fake follower, spammer and other forms of automated accounts may be combining to increase fear in the community about topics related to the Russia/Ukraine war. This observation has been hypothesised by many authors such as [106, 110], but a detailed analysis has been lacking and may be of concern for many governments and defence organisations.

Figure 4.6 further shows that fake follower, self-declared, spammer and other bot types also Granger-cause an increase in online discussion around motion. In Figure 4.7b, we see a number of motion related words that are potentially associated with staying or fleeing the country. Combining this with increases in angst suggests a relationship between the activity of these types of bots and discussions of humanitarian movement within Ukraine. Druziuk [50] noted that bots have allowed "Ukrainians to report to the government when they spot Russian troops", but the usage of bots to influence people on staying/leaving the country is something not observed before. According to an article by Osborne [110], over 100,000 bot accounts were found to have "inspire panic among Ukrainian citizens and destabilize the socio-political situation in various regions". The same article suggests that the Security Service of Ukraine (SBU) had destroyed these bot farms and accused Russia of operating the farms for conducting a "large-scale information sabotage". If this is the case, then it is likely the increases in angst are due to these bot farms frightening Ukrainian citizens with the intention to influence their opinion on staying or fleeing the country. However, it is difficult to deduce exactly why Russia may have these intentions.

We additionally observe the self-declared bots to Granger-cause an increase in discussions surrounding time. In Figure 4.7e we see that words related to 'time' appear to be words like 'now', 'ever', 'today' and 'time', and we observe these effects to be most prominent after 5 hours. More discussions surrounding time indicate that accounts may be discussing things which are not related to the present – a sign of worry [66]. Consequently, the increases in angst (which are most prominent after 7 hours) may be a result of worries about the future. However, we can only hypothesize these effects as we did not consider the effect of anxiety on worry directly.

In Figure 4.6 the most prolific lag is mostly consistent for a given LIWC category, but varies greatly for bot type. Hence, the time which bots effect a given discussion on the war depends mainly on the topic of discussion and not on the type of bot. For instance, we observe that Fake Follower, Self-Declared, Spammer and Other bots all most prolifically effect discussions of work after five hours. To further examine the effects of the lag on discussion of different LIWC categories, we plot cross correlations in Figure 4.8. These plots represent the cross correlation between Self-Declared bot proportions and a number of significant LIWC categories (in Figure 4.6) over 48 hours. Note that cross correlations are the correlations between two variables at different lags.

The direction of the effect for each LIWC variable in Figure 4.8 is consistent with Figure 4.6, further validating our results. This direction is consistent for all significant lags, justifying our decision to choose the largest parameter in the regression model as an indication of direction in Figure 4.6. For some LIWC categories the effects of Self-Declared bots linger over many lags but for others the effects diminish relatively quickly.

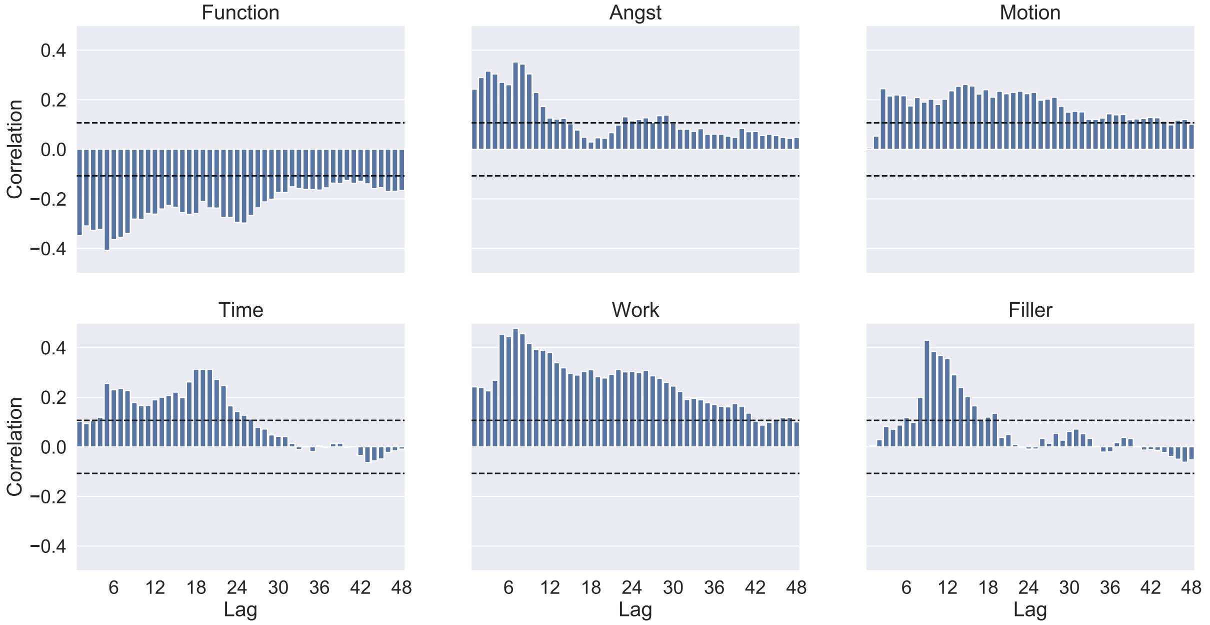


Figure 4.8: Lagged cross correlations between the hourly Self Declared bot proportions and the significant hourly LIWC Category proportions (significance is determined from the results in Figure 4.6). We consider 48 hours/lags for each of these plots and represent the significance threshold using a horizontal dotted line. This plot can be used to examine the extent to which the bots drive changes in online discussion and how long these effects can persist for.

For instance, the effects on the work category and the function category are significant for lags almost up to 48 hours, whereas the effects on the angst and filler categories diminish within 24 hours. This indicates that different conversation topics can persist on Twitter for some time, whereas others diminish faster. Nonetheless, the length of time that a conversation topic persists seems to be dependent on the topic itself, rather than the type of bot activity. Next we wanted to observe the significant effects from the self-declared bots on the various LIWC categories. To do this, we considered the time series of these different LIWC categories relative to the self-declared bots as seen in Figure 4.9. We include several significant events on the figure, as we have done with previous time-series plots.

As we have previously observed in Figure 4.1, the self-declared bot probabilities have a significant spike on the 4th of March when Russia captured the Zaporizhzhia nuclear power plant. Shortly after this we observe the spike in the self-declared bot activity and consequently changes in the discussion topics of several LIWC categories. As a result, the time series further demonstrate the effects of the self-declared bots, supporting our previous remarks relating to the effects of bots on the online discussion. While the effects

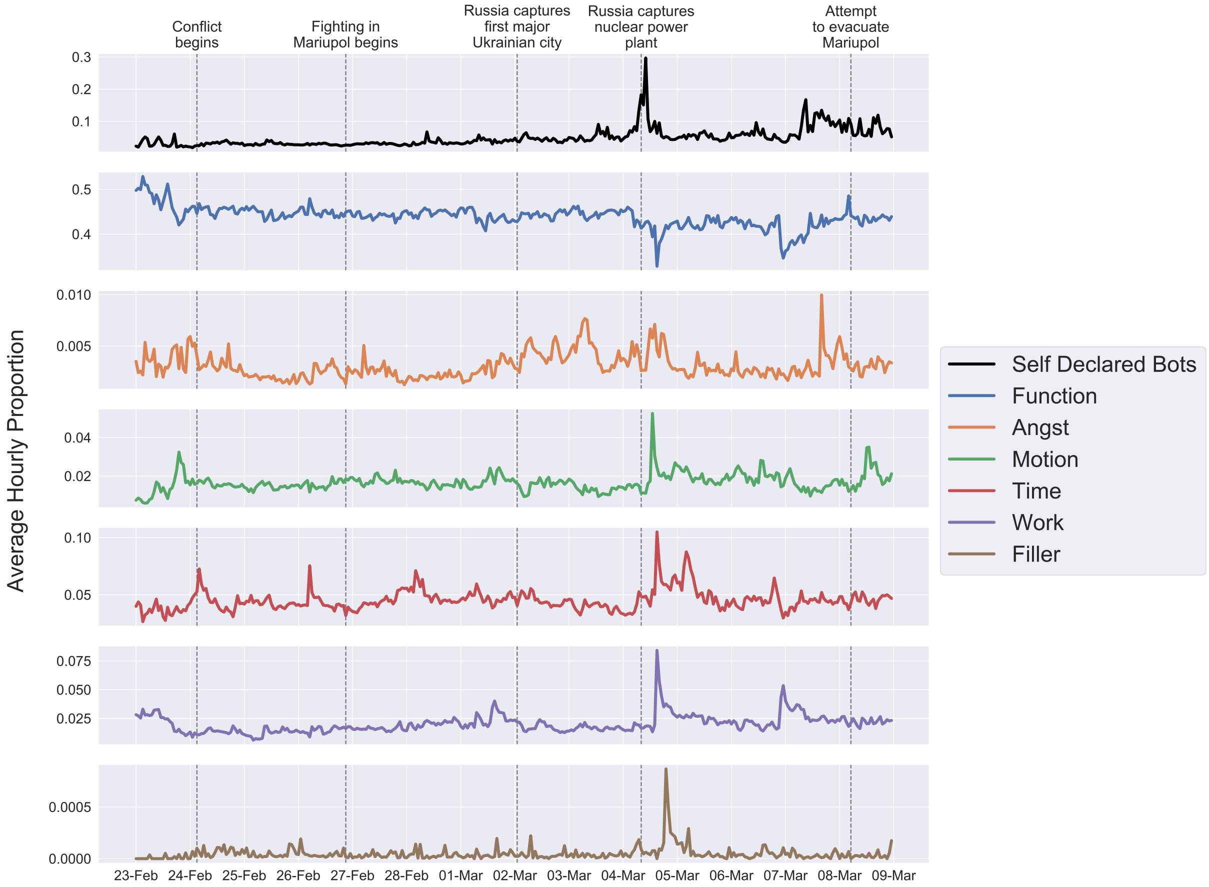


Figure 4.9: Time series plots of the self-declared bot activity as well as six different LIWC categories. We choose these six LIWC categories because they are significantly effected by the activity of the self-declared bots (see Figure 4.6). We include several significant events on the figure, as we have done with previous time-series plots.

of the self-declared bots are most observable around the 4th of March, we acknowledge that these effects are present over the whole fortnight of data through utilising the Granger Causality Test (see Section 2.4.2 for details).

Summary

The work in this chapter extends existing techniques to understand how bot-like accounts spread misinformation/disinformation on Twitter and measures the effect of these malicious campaigns. The main contributions of this chapter include a dataset[3] of ap-

[3] Dataset available at `https://figshare.com/articles/dataset/Tweet_IDs_Botometer_results/20486910`.

proximately 5.2 million posts created by Twitter users who participated in discussions around the Russian Invasion of Ukraine [158]. Using this dataset, we identify bots which contribute to the online discussion related to the Russia/Ukraine war. We consider the activity of bots over the fortnight after Russia's invasion of Ukraine and how this activity aligns with significant events over this time period. We consider the distributions of different bot types contributing to the discussion by viewing box plots and probability density functions of these probabilities. We provide an analysis of the effect which bot activity has on emotions in online discussions of the conflict. We find that bots significantly increase discussions of the LIWC categories: Angst, Friend, Motion, Time, Work and Filler. The strongest relationship is between Self Declared bot activity and words in the 'Work' category (with $p = 3.803 \times 10^{-18}$), which includes words relating to governance structures like 'president' and 'government'. More generally, we create a statistical framework which can be applied to measure the influence of a group of users in a network. While these techniques have been applied to the context of the Russia/Ukraine war, they may also be applied more generally in political campaigns, dis/misinformation campaigns or any online advertisement campaign.

While we provide a comprehensive analysis of the dataset we collected, we also acknowledge this study is the first piece of published work in this area on the 2022 Russian/Ukraine war. Hence, there is a large scope for future research in this space, using our dataset as well as new datasets. Since our study has fundamentally relied upon the accuracy of Botometer, one avenue of future work could consider a similar analysis but using different bot detection methods. This would enable the validity/accuracy of Botometer to be explored, and it would allow other researchers to either confirm or deny our findings in this paper. Other areas of future research could explore information contained on the network of interactions between users recorded in the dataset. This may include looking at the network of retweets, quote tweets, replies or a combination of the three. Doing this would allow researchers to apply a network science approach to our dataset, similar to authors such as Barabási and Pósfai [11], Newman [104]. Examining coordination in these networks would allow researchers to quantify the impact of coordinated activity in the social network structure and further investigate its influence on social media users. Other research could consider the timing lags between the posts of various accounts, and use this to detect coordinated activity between users in our dataset. If the distributions of the timing lags were heavy-tailed and differed between account types, this would suggest differences in coordinated activity signatures – something previously explored by Mathews *et al.* [92]. Finally, it would also be important to explore diverse ways of classifying the national lean of authors based on their published Twitter content. Another way of classifying this national lean would be to only classify accounts as pro-Ukraine or pro-Russian if all their hashtags were in support of either Ukraine or Russia, respectively. Using this method, the balanced account group would be much larger and would contain all users that have used hashtags supporting both sides, potentially leading to different results.

In the next chapter we will summarise the research performed in all previous chapters. This will include a road-map outlining the results of each chapter in the context of our original research ambitions. We will then consider a placement of our work in the context of the original research question. Finally, we will discuss any future research in this field and how our results may be used to support this future research.

Chapter 5

Conclusion

Our work aimed to develop statistical frameworks which underpin two key vulnerabilities related to the information environment created by social media. These two potential vulnerabilities are: (i) the huge amount of individual-level data that is present on these applications, and (ii) the underlying dialogic transmission system; with many sources of information and many receivers. It is important to understand how these aspects of social media can be weaponised by individuals and groups of individuals from a psychological perspective. In our analysis, we considered the psychological impact of both of these problems, where we considered the personality types of individual accounts and the cognitive influence of groups of online accounts. In essence, we consider the personal information that can be learnt about accounts on an individual level and we consider whether groups of automated online accounts can influence a large number of human accounts.

In Chapter 2, we provided a background of work to contextualise the research and define our research objectives. Initially, we provided an overview of two different personality models: the OCEAN personality model and the Myers-Briggs personality model. We distinguished between the psychological properties of the two models: the OCEAN model is a five factor model which was formulated using a statistical approach and the Myers-Briggs model is a four factor model which was created using a theory driven approach. While we used the Myers-Briggs personality model in our analysis, much of the literature in this field to date has used the OCEAN model and so highlighting the features of this model was necessary. Next we outlined the Natural Language Processing (NLP) tools we utilised in our analysis, this included: Linguistic Inquiry and Word Count (LIWC; pronounced "Luke") [113], Valence Aware Dictionary for Sentiment Reasoning (VADER) [74], Bidirectional Encoder Representations from Transformers (BERT) [41], and Botometer [163], a supervised machine learning classifier which distinguishes bot-like and human-like accounts on Twitter. Next we provided a mathematical background of our work, consisting of: the binary models used during personality profiling, the statistical methods for hypothesis testing and any data manipulation methods used in our analysis. Finally, we performed a literature review of the research to date which has focused on either online personality profiling or the detection/influence of automated online accounts (bots). This allowed us to identify critical gaps in the research to date and consequently define research objectives to address this. For instance, we found that most of the personality models to date have been created with the purpose of being highly predictive. This identified a requirement for more interpretable models in this area, allowing us to interpret model parameters in the context of the original problem. Moreover, we found that a lot of previously existing labelled datasets in this field have been highly unbalanced and there is limited research which focuses on these problems. Consequently, we performed various weighting/sampling techniques in our work to address this. On the topic of bot detection/influence, we discovered that most of the research focuses on the detection of bots, and their influence is only hypothesised a lot of the time. Furthermore, there is no work (to our knowledge) which addresses their influence during the 2022 Russia/Ukraine war. This highlighted the importance of discovering their influence, and how this shapes online discussion surrounding the war.

In Chapter 3, we determined how informative someone's online digital footprint was in predicting their Myers-Briggs personality type. We collected a dataset of Twitter accounts with labelled Myers-Briggs personality types. We did this by querying for accounts which had self-reported their personality types on Twitter. Our data collection techniques were novel and allowed us to create the largest labelled dataset of Myers-Briggs personality types on Twitter (to our knowledge) at low cost. We collected a number of different features for these accounts which included social metadata features as well as linguistic features – including LIWC, VADER, BERT and Botometer features. We then preprocessed the data as well as introduced an inclusion-exclusion criteria for the accounts. Next we performed an exploratory data analysis (EDA) on the dataset, where we found that our dataset contained some biases related to strictly focusing on Twitter accounts. However, we acknowledge that our models can only be applied to Twitter accounts and new labelled datasets from other social media applications would be required to profile characteristics on these sites. Moreover, we found that some of the Myers-Briggs dichotomies are very unbalanced in our data, so we performed five different weighting/sampling techniques prior to fitting logistic regression, naive Bayes, support vector machine and random forest models. Because of these imbalances, we discovered researchers need to be vigilant of the metrics they choose to use when assessing model performance. For instance, reporting accuracies can be misleading; because models can utilise the class imbalances to predict the majority classes. Consequently, we compared our accuracies to a majority class classifier, and additionally reported more suitable metrics such as AUC, F1, precision and recall. However, we found that most scholars in this area had not previously acknowledged these class imbalances when modelling similar datasets and yet, still reported accuracies for their models with no reference to these imbalances [117, 15]. Next we compared the results of the logistic regression models with a number of other machine learning models, including: naive Bayes, support vector machines and random forests. We found that the support vector machine models generally perform the best, where we achieved an accuracy of 20.95% for accurately predicting someones complete Myers-Briggs personality type – only 5.64% larger than the majority class classifier. Hence, we deduced modelling personality types was difficult on Twitter and also likely less trivial that it appeared for other scholars considering different platforms (due to similar class imbalances). Moreover, we discovered that synthetic minority oversampling technique (SMOTE) performs very poorly when combined with naive Bayes. Hence, we performed a low-dimensional example of SMOTE with naive Bayes to outline a discrepancy in combining both these techniques on certain data. Lastly, we formulated a statistical framework for analysing the importance of different features in our models. In particular, we found that some groups of features were more informative when modelling the Intuitive/Sensory dichotomy ($p = 0.032$) and the Thinking/Feeling dichotomy ($p = 0.019$). Altogether, this chapter quantified how an environment of huge amounts of data, like social media, can be utilised to profile personal characteristics about individuals at large scale.

In the Chapter 4, we discovered how influential bot accounts were in the online discussion of the 2022 Russia/Ukraine war on Twitter. We did this by collecting a dataset of tweets, quotes, retweets and replies related to the war over the first two weeks since Russia invaded Ukraine. We discovered the most trending hashtags relevant to the war and queried those which were in support of either Russia/Putin or Ukraine/Zelenskyy. We used these hashtags to calculate the national 'lean' for the accounts, outlining whether an account was in support of Russia/Putin, Ukraine/Zelenskyy or both. Moreover, we calculated the Botometer scores for a random sample of these accounts. We then considered a time series analysis of bot activity and observed how it aligned with the frequency of the hashtags as well as a number of significant events which occurred over the first two weeks of the war. As part of this, we found there to be significant spikes in bot activity when Russia captured its first major Ukrainian city (Kherson) and when Russia captured the Zaporizhzhia nuclear power plant. Next we observed a normalised distribution of bot probabilities based on the national 'lean' of accounts. Both distributions appeared bimodal, however the distribution of pro-Ukrainian accounts displayed heavier tails than the distribution of pro-Russian accounts. Hence, it appears that pro-Russian bots may be harder to detect than pro-Ukrainian bots – perhaps because the authorities creating these bots are more equipped to hide their bot-like behaviour. We then considered the effects of bot accounts on the overall discussion surrounding the war. We did this by formulating a statistical framework which enabled us to discover the effects of bot activity on several linguistic features of the discussion. We observed that bots strongly increased discussions of work/governance ($p = 3.80 \times 10^{-18}$) with the most prominent effects after five hours. Moreover, we discovered bots also increased angst in the online discourse ($p = 2.45 \times 10^{-4}$) as well as discussions of motion ($p = 7.93 \times 10^{-10}$) with the most prominent effects after seven and three hours, respectively. Since discussions of motion were most often involved with staying/fleeing the country, we deduced that it is likely bots were influencing peoples' decision to flee their country or not. Consequently, this chapter allowed us to discover how social media can be weaponised by governments in a modern-day conflict to influence public discussion at a large scale. We discovered the underlying dialogic transmission system was being weaponised by the creators of automated online accounts to shape the discussions on the Twittersphere surrounding the Russia/Ukraine War.

Social media creates an information environment with two important aspects that are unique to this type of medium. Firstly, there are large amounts of individual-level data that each user provides, and secondly, there is an underlying dialogic transmission system; where there are many sources of information and many receivers. These two aspects of social media make it very different from traditional media, such as newspapers, TV and radio, where there is very limited individual-level data and very few sources of information. However, these aspects of social media have been suggested to make these sites more vulnerable to weaponisation by various companies and governments [162, 111, 143]. In our work we developed statistical frameworks which underpin the vulnerabilities in these unique aspects of social media. We separated our analysis into two chapters which addressed each of these aspects individually.

Firstly, we discovered that someone's online digital footprint was not that informative of personal information about online accounts, like their personality types. In essence, this work exposes that companies like Cambridge Analytica may not have been able to expose personal attributes about online accounts with a high degree of accuracy. However, scholars such as Sumpter [142] have suggested that Cambridge Analytica did not have highly accurate models to perform their targeted political advertisement campaigns, but rather had models with similar performance to what we observed and performed these political advertisement campaigns at a very large scale so the effects were noticeable. For example, an improvement of 5.64% above a majority class classifier, like we observed in our models, would account for 28,200 more correctly classified accounts in a large scale political advertisement campaign of 500,000 accounts. As a result, our models would likely be capable of performing political advertisement campaigns to the same extent as Cambridge Analytica, highlighting the need for strict regulations regarding who gets access to such personal data.

We then aimed to describe how agencies are utilising automated accounts (bots) to weaponise the dialogic transmission system which underpins social media. Agencies are suggested to do this by spreading online misinformation/disinformation which promotes their own motivations and destabalises their opponents objectives. We developed techniques which measured the influence of bots in the Russia/Ukraine war, but these techniques can be more widely used to quantify and contextualise the influence of any type of online campaign. Most concerningly, we showed these bots drove increases in angst across the wider online community – something which should be of concern for anyone using these platforms as well as governments which aim to maintain a safe online community.

Our findings highlight the need for defence technologies which can distinguish between human-like and bot-like accounts as well as defence technologies which can distinguish between true and false information.

Our research has quantified how the information environment underpinning social media can be weaponised. This has highlighted the need for policies protecting people's personal data as well as defence technologies which can prevent the ability for external agencies to perform unfavourable public influence at large scales. While our analysis was comprehensive, we acknowledge there is a need for future work in this area. Firstly, there is a need for future work addressing the extent to which someone's online digital footprint can be used to profile personal characteristics about accounts. Our analysis only considered a small aspect of this problem; four different machine learning models on Myers-Briggs personality types. Since our dataset contained a large number of accounts and features, there is certainly the potential for a deep learning model to be trained on the same data. recurrent neural networks, transformer models, Long Short Term Memory (LSTM) and a number of other black-box type machine learning models could all be useful candidates. While these models would not give the same interpretability as some of our models, they may achieve better predictive capabilities on our data. One other area of future work may consider different linguistic features in these models. The linguistic features we collected are a very small sample of features which were chosen because they have achieved high accuracies in a number of other models [6, 12, 144]. However, there are a large number of different NLP tools such as Word2Vec, ELMo and GloVe which could be utilised for this type of task. It is also important to note that the BERTweet model used in our analysis was pre-trained for two specific language tasks: language modelling and next sentence prediction. Hence, another extension of the work done here would be to fine-tune BERTweet for Myers-Briggs personality prediction, and this would likely improve results. Other future work should also consider using different data collection methods. One limitation of our dataset is that we only have access to the classification of each personality dimension, when in reality these dimensions exist on a numerical scale [147]. For instance, two users may both be extroverted but one user may be far more extroverted than the other. While performing questionnaires are expensive, it would enable us to obtain these personality dimensions on a numerical scale, and we would expect it to improve the performance of our models. Another obvious extensions of our work would be to use the OCEAN personality model rather than the Myers-Briggs personality model. If we utilised questionnaires to collect our data, we would have the luxury to choose which personality model we use and this would potentially achieve different results. However, there were obvious reasons why we chose our data collection techniques: they were novel and enabled us to achieve a labelled dataset without these long, expensive and cumbersome questionnaires.

There is also a need for future work which focuses on measuring and quantifying online influence operations – this may be by bots, misinformation, disinformation or various other forms of malicious activity. While we provided a comprehensive analysis of the dataset we collected on the Russia/Ukraine war, we also recognise this conflict is ongoing and there is a need for future research in this area. We acknowledge that our approach has fundamentally relied on the accuracy of Botometer, so future work should focus on using/developing various other bot detection methods. This would enable researchers to perform a similar approach to us and potentially validate our results using different methods. It would also enable scholars to explore the accuracy of Botometer – something which has often been questioned by academics and even described as a "tool that everybody can use to produce pseudoscience" [96]. These bot detection algorithms are claimed to often overestimate the probability that online accounts are bots [96, 55]. Gallwitz and Kreil [55] showed a large proportion of accounts with high bot scores were operated by human users and suggested that these accounts were "using Twitter in an inconspicuous and unremarkable fashion without the slightest traces of automation". As a result, there is certainly a need for further work which addresses the ability for bot detection algorithms to distinguish between automated online accounts and malicious human actors. Another area of future research may consider using more network science based approaches. This could include looking at networks of interactions between users recorded in our dataset. Examining coordination in these networks would potentially allow researchers to quantify the impact of coordinated activity in the social network structure and further investigate this type of activity on social media users. Other future work may consider the timings which accounts are posting and use this to determine coordinated activity. For instance, if the distributions of timing lags differed between account types and were heavy-tailed, this would suggest diverse signatures of coordinated activity and further support findings by authors such as Matthews [93]. We could explore different ways of classifying the national lean of accounts based on what they post on Twitter. For instance, authors could classify accounts to be pro-Russian or pro-Ukrainian if every hashtag an account shares is in support of either Russia/Putin or Ukraine/Zelenskyy, respectively. So the balanced account group would contain any user sharing a mixture of hashtags supporting each side, regardless of the frequency of pro-Putin/pro-Ukraine hashtags used. Using this approach may lead to different results and would allow researchers to explore the sensitivity of our results to changes in these methods. As a whole, the Russia/Ukraine war has sparked large volumes of online activity discussing the war and this type of activity provides a unique opportunity for researchers to quantify/understand the role of the information environment during a modern-day conflict.

Appendix A

Personality Profiling Model Results

In this section, we include the results for the naive bayes, support vector machine and random forests classification models performed on the personality dataset. These results supplement the material included in Chapter 3.

A.1 Naive Bayes Classifier

We implemented the naive bayes classifier on each of the four dichotomies in the personality dataset – an overview of the naive bayes classifier is provided in Section 2.4.1. We perform four independent classifiers on each of the dichotomies. Similar to the logistic regression models, we perform different weighting/sampling techniques on our data due to the unbalanced nature of the dichotomies. We use the same sampling techniques as we developed in Section 3.2.1, however the weighted model is performed slightly differently to the weighted logistic regression model. We provide a detailed explanation of the weighted model in Section 3.2.2.

We perform the upsampled, SMOTE and downsampled models until the class balance of each dichotomy in the training set is equal. For the model combining SMOTE and downsampling, we sampled the classes so they had the same ratio of observations as the 'SMOTE + Downsampled' logistic regression models. Recall these were ratios of 90%, 80%, 85% and 100% for the Extroverted/Introverted, Intuitive/Sensory, Thinking/Feeling and Judging/Perceiving dichotomies, respectively. Note that we used the same ratios as the afore discussed models for consistency and so we can compare the results from the two models. In each case, we only perform these sampling techniques on the training set prior to fitting every model.

In each model, we used the first 200 principal components of the features and these components were determined using only the training data. While naive bayes models are very quick to train and test, we did this so our models were comparable with results in the previous section. In each case, we perform cross validation with ten splits. First, we consider the results of the five classifiers by calculating their accuracy, AUC, F1, precision and recall. The results are displayed in Figure A.1.

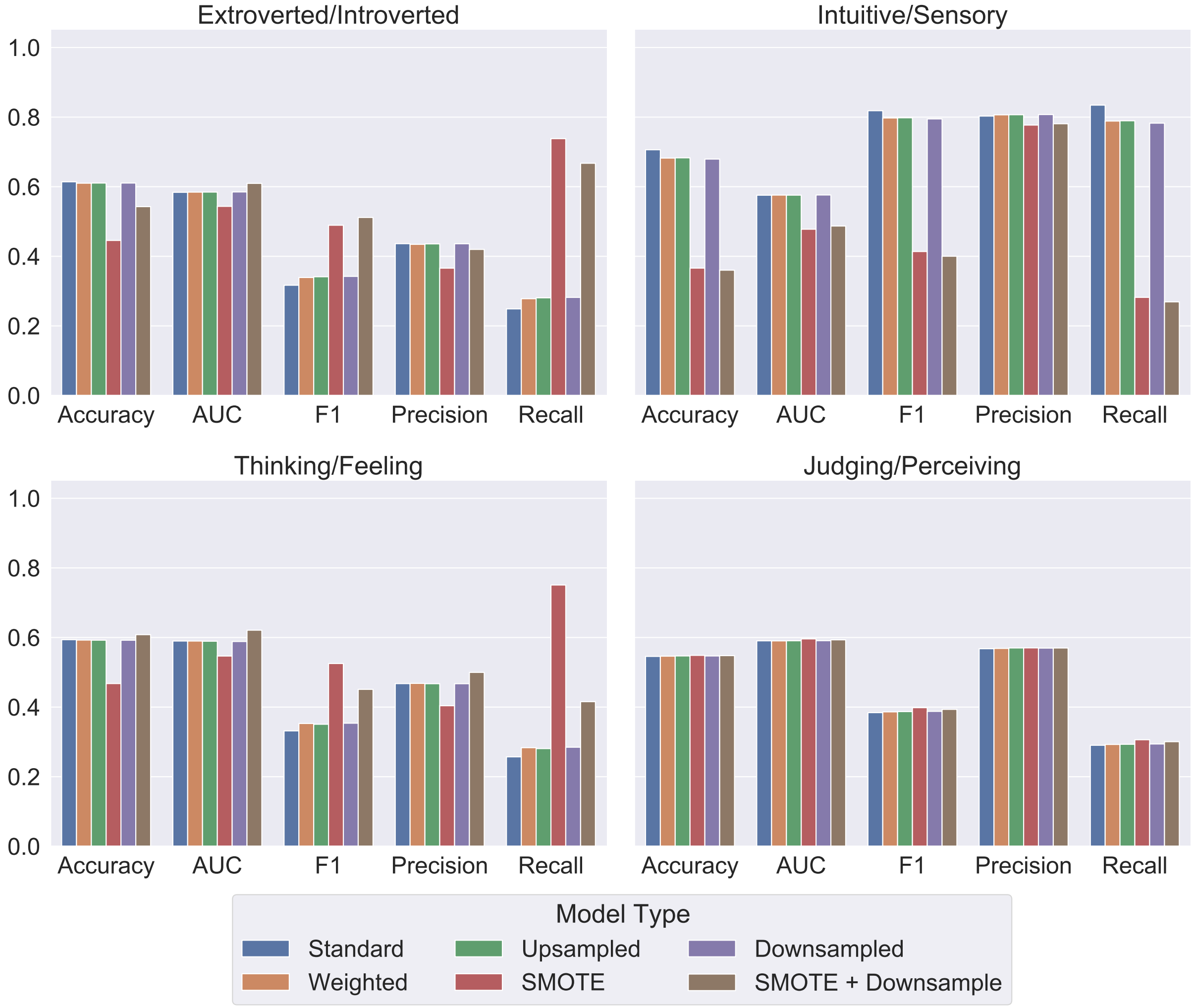


Figure A.1: Summary of metrics from the naive bayes models of the four independent Myers-Briggs dichotomies. These metrics include the Accuracy, Area Under the Curve (AUC), F1, Precision and Recall Scores. For each model, these metrics were calculated based on a ten-fold cross validation

Next we wanted to assess the performance of our models more holistically by combining the results for each model across the independent dichotomies. We do this by looking at the Receiver Operating Characteristic (ROC) curve of our models performance on the testing data. Performing the models independently for each dichotomy means that there is six models of the four different dichotomies, a total of 24 different naive bayes models.

Like the logistic regression model in Section 3.2.1, we macro-average and micro-average the TPR/FPR for the ROC curves using equations 3.2 and 3.3, respectively. We performed these calculations at each threshold for every naive bayes model type and displayed the results in an ROC curve, seen in Figure A.2. We also include the Area Under the Curve (AUC) metric for each of the ROC curves.

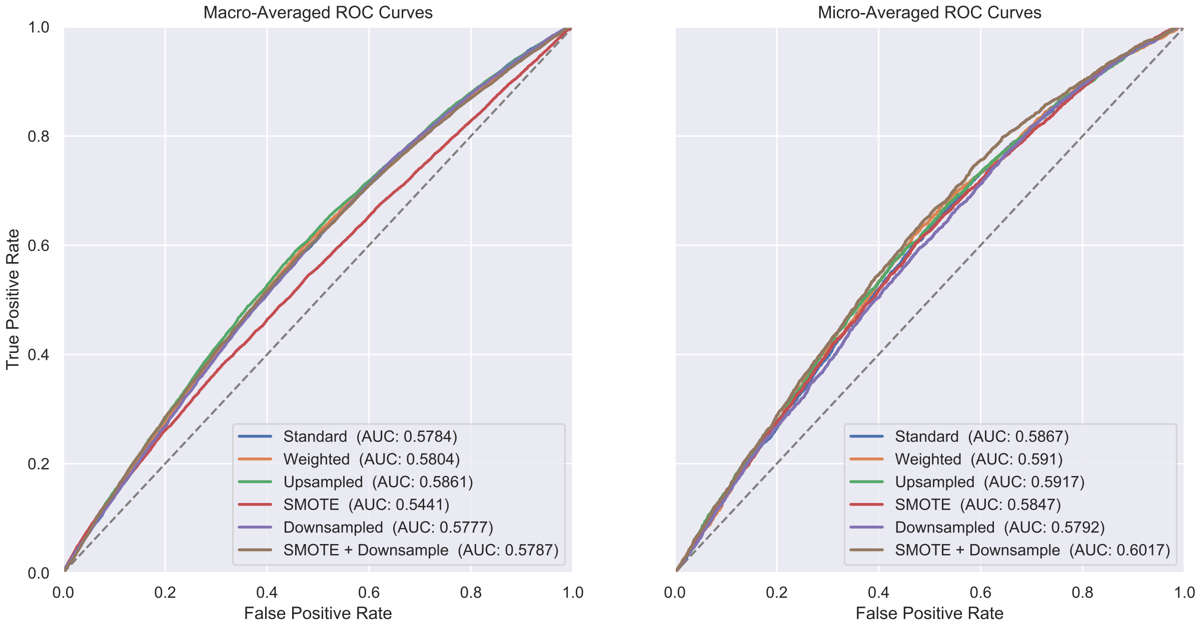


Figure A.2: Macro-averaged (left) and Micro-averaged (right) Receiver Operating Characteristic (ROC) curves for each naive bayes model. Note that we macro-average and micro-average the results from the independent models for each dichotomy using Equations 3.2 and 3.3, respectively. We include the Area Under the Curve (AUC) metric for each of the ROC curves as well.

For comparability to the logistic regression models presented in Section 3.2.1, we also present the accuracies of the naive bayes models. This is because the accuracy is the most easily understood metric which is directly informative of a models performance on new data. The accuracy is also the most commonly reported metric by other researchers in this field, making the results comparable across various papers and models. As a result, the accuracies for each of the naive bayes models are reported in Table A.1. We report four types of accuracy depending on the number of accurately predicted dichotomies in each model. These are the proportion of users where we accurately predict all four dichotomies, at least 3 dichotomies, at least 2 dichotomies and at least one dichotomy. However to ensure our accuracies are not misleading, we report the results for a model where we sample personality types randomly and a model where we simply select the majority class. We report the theoretical accuracies for the random classifier and we determine the accuracies for the majority classifier using the proportion of types in Figure 3.1. Note that the best models for each metric are bolded.

	Accurately Predicted Dichotomies			
Model	4	≥ 3	≥ 2	≥ 1
Standard	14.20	49.17	81.91	97.40
Weighted	13.76	48.02	80.82	97.18
Upsampled	13.75	48.06	80.82	97.18
SMOTE	6.20	25.65	60.63	90.62
Downsampled	13.60	47.87	80.86	97.20
SMOTE + Downsample	6.41	33.78	69.89	93.25
Random Classifier	6.250	31.25	68.75	93.75
Majority Class	**15.31**	**54.54**	**87.20**	**98.28**

Table A.1: Reported accuracies for the naive bayes models. For each model, we include the accuracy of correctly predicting all four dichotomies, at least three dichotomies, at least two dichotomies and at least one dichotomy. We report the theoretical accuracies of a random classifier and we determine the accuracies of a majority class classifier using the proportion of types observed in Figure 3.1.

A.2 Support Vector Machine Classifier

We implemented the support vector machine classifier on each of the four dichotomies in the personality dataset – an overview of the support vector machine classifier is provided in Section 2.4.1. We perform four independent classifiers on each of the dichotomies. We use the same sampling techniques as we developed in Section 3.2.1, however the weighted model is performed slightly differently to the afore mentioned models. We provide a detailed explanation of the weighted model in Section 3.2.2.

We perform the upsampled, SMOTE and downsampled models until the class balance of each dichotomy in the training set is equal. For the model combining SMOTE and downsampling, we sampled the classes so they had the same ratio of observations as the previous models where we have combined SMOTE and downsampling. Note that we used the same ratios as the afore discussed models for consistency and so we can compare the results from the two models. In each case, we only perform these sampling techniques on the training set prior to fitting every model.

In each model, we used the first 200 principal components of the features and these components were determined using only the training data. Using the first 200 principal

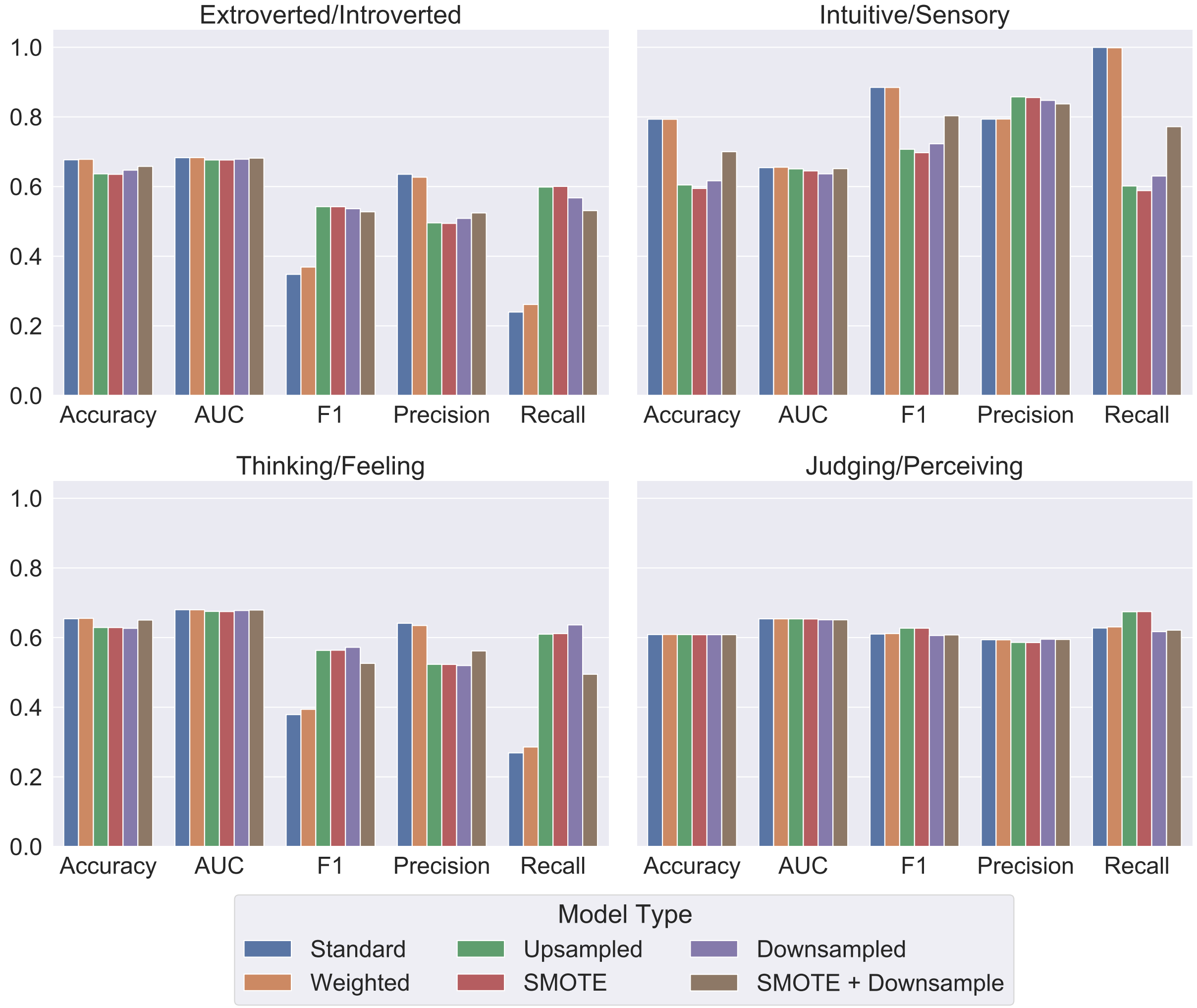


Figure A.3: Summary of metrics from the support vector machine models of the four independent Myers-Briggs dichotomies. These metrics include the Accuracy, Area Under the Curve (AUC), F1, Precision and Recall Scores. For each model, these metrics were calculated based on a ten-fold cross validation

components was necessary to reduce the computational time for fitting the models. The computational time of fitting the support vector machine model used in our analysis scales between $\mathcal{O}\left(n_{\text{features}} \times n^2_{\text{samples}}\right)$ and $\mathcal{O}\left(n_{\text{features}} \times n^3_{\text{samples}}\right)$. Hence, using PCA to reduce the dimension of our feature space significantly improved the computational time. When fitting each model, we perform cross validation with ten splits on the dataset. First, we consider the results of the five classifiers by calculating their accuracy, AUC, F1, precision and recall. The results are displayed in Figure A.3.

Next we wanted to assess the performance of our models more holistically by combining the results for each model across the independent dichotomies. We do this by looking at the Receiver Operating Characteristic (ROC) curve of our models performance on the testing data. Like the afore mentioned models, we macro-average and micro-average the TPR/FPR for the ROC curves using equations 3.2 and 3.3, respectively. We performed these calculations at each threshold for every support vector machine model type and displayed the results in an ROC curve, seen in Figure A.4. We also include the Area Under the Curve (AUC) metric for each of the ROC curves.

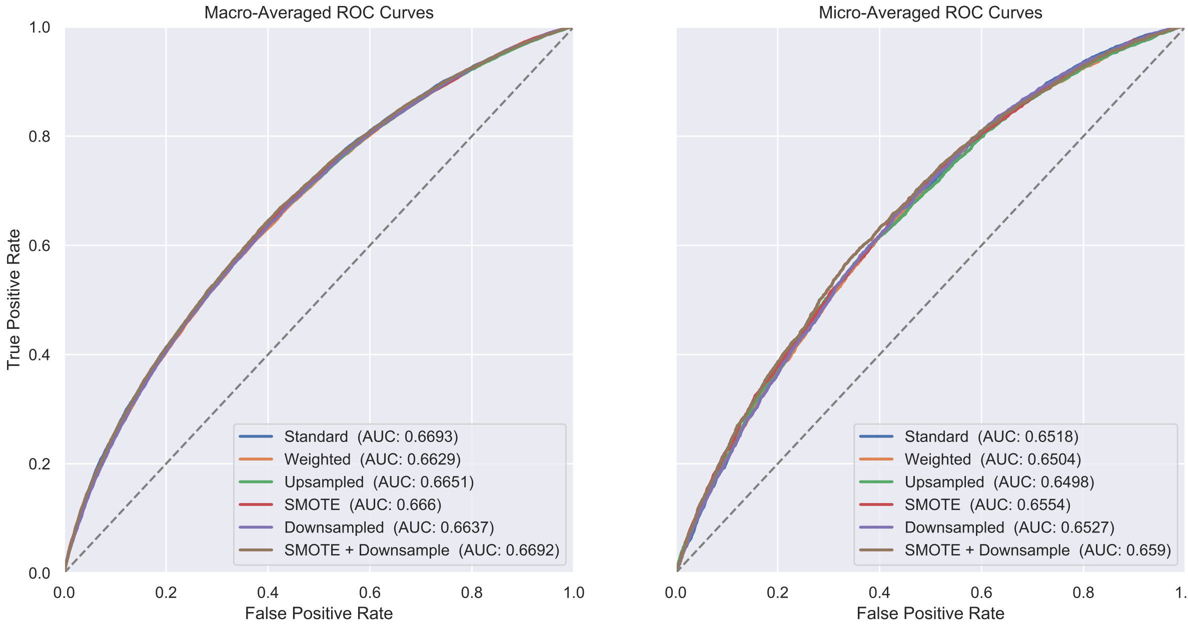


Figure A.4: Macro-averaged (left) and Micro-averaged (right) Receiver Operating Characteristic (ROC) curves for each support vector machine model. Note that we macro-average and micro-average the results from the independent models for each dichotomy using Equations 3.2 and 3.3, respectively. We include the Area Under the Curve (AUC) metric for each of the ROC curves as well.

For comparability to the previously discussed models, we also present the accuracies of the support vector machine models. This enables the results to be comparable across various papers and models. As a result, the accuracies for each of the support vector machine models are reported in Table A.1. We report four types of accuracy depending on the number of accurately predicted dichotomies in each model. These are the proportion of users where we accurately predict all four dichotomies, at least 3 dichotomies, at least 2 dichotomies and at least one dichotomy. However to ensure our accuracies are not misleading, we report the results for a model where we sample personality types randomly and a model where we simply select the majority class. We report the theoretical accuracies for the random classifier and we determine the accuracies for the majority classifier using the proportion of types in Figure 3.1. Note that the best models for each metric are bolded.

	Accurately Predicted Dichotomies			
Model	4	≥ 3	≥ 2	≥ 1
Standard	**20.95**	60.25	89.64	**98.90**
Weighted	20.91	**60.54**	**89.71**	**98.90**
Upsampled	13.49	49.14	83.05	97.59
SMOTE	13.56	48.61	82.54	97.61
Downsampled	13.61	49.18	83.13	97.84
SMOTE + Downsample	17.00	54.26	85.81	98.19
Random Classifier	6.250	31.25	68.75	93.75
Majority Class	15.31	54.54	87.20	98.28

Table A.2: Reported accuracies for the support vector machine models. For each model, we include the accuracy of correctly predicting all four dichotomies, at least three dichotomies, at least two dichotomies and at least one dichotomy. We report the theoretical accuracies of a random classifier and we determine the accuracies of a majority class classifier using the proportion of types observed in Figure 3.1.

A.3 Random Forest Classifier

We implemented the random forests classifier on each of the four dichotomies in the personality dataset – an overview of the random forests classifier is provided in Section 2.4.1. We perform four independent classifiers on each of the dichotomies. We use the same sampling techniques as we developed in Section 3.2.1, however the weighted model is performed slightly differently to the afore mentioned models and this is explained in Section 3.2.2.

We perform the upsampled, SMOTE and downsampled models until the class balance of each dichotomy in the training set is equal. For the model combining SMOTE and downsampling, we sampled the classes so they had the same ratio of observations as the previous models where we have combined SMOTE and downsampling. Note that we used the same ratios as the afore discussed models for consistency and so we can compare the results from the two models. In each case, we only perform these sampling techniques on the training set prior to fitting every model.

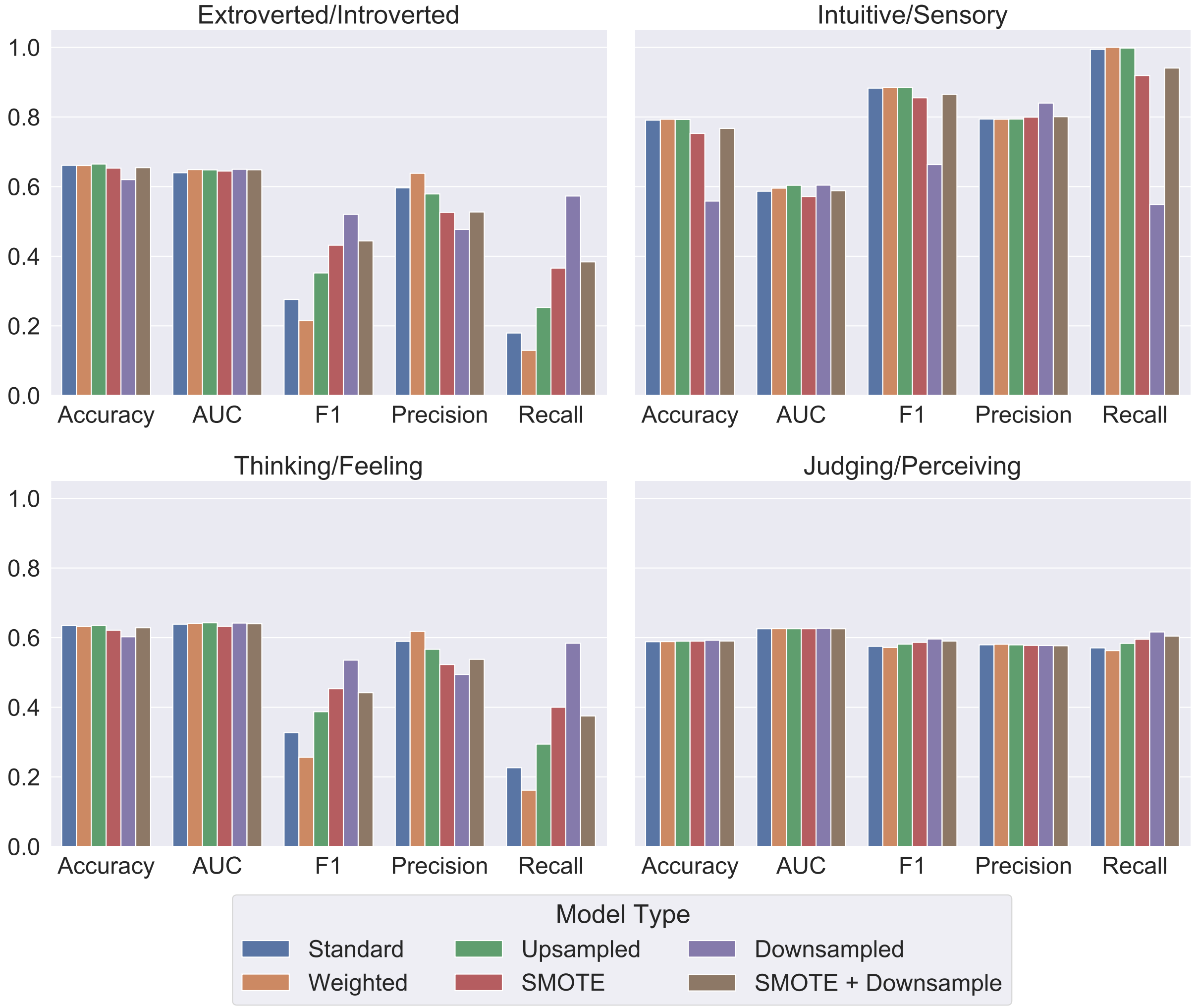


Figure A.5: Summary of metrics from the random forests models of the four independent Myers-Briggs dichotomies. These metrics include the Accuracy, Area Under the Curve (AUC), F1, Precision and Recall Scores. For each model, these metrics were calculated based on a ten-fold cross validation

In each model, we used the first 200 principal components of the features and these components were determined using only the training data. This significantly reduced the computational time for fitting each of the models – this is because the computational complexity of training random forests is $\mathcal{O}\left(n_{\text{features}} \times n_{\text{samples}} \times \log n_{\text{samples}}\right)$. When fitting each model, we perform cross validation with ten splits on the dataset. First, we consider the results of the five classifiers by calculating their accuracy, AUC, F1, precision and recall. The results are displayed in Figure A.5.

Next we assessed the performance of our models by combining the results for each model across the independent dichotomies. We do this by looking at the Receiver Operating Characteristic (ROC) curve of our models performance on the testing data. Like the afore mentioned models, we macro-average and micro-average the TPR/FPR for the ROC curves using equations 3.2 and 3.3, respectively. We performed these calculations at each threshold for every random forests model type and displayed the results in an ROC curve, seen in Figure A.6. We also include the Area Under the Curve (AUC) metric for each of the ROC curves.

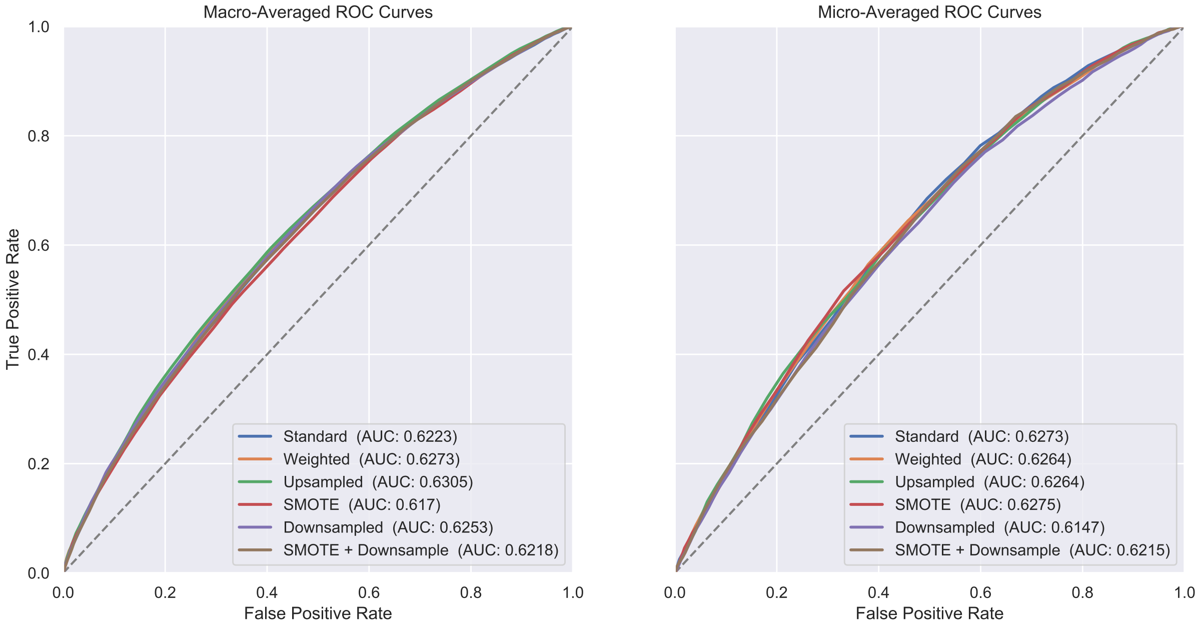


Figure A.6: Macro-averaged (left) and Micro-averaged (right) Receiver Operating Characteristic (ROC) curves for each random forests model. Note that we macro-average and micro-average the results from the independent models for each dichotomy using Equations 3.2 and 3.3, respectively. We include the Area Under the Curve (AUC) metric for each of the ROC curves as well.

For comparability to the previously discussed models, we also present the accuracies of the random forests models. This enables the results to be comparable across our previous models and personality models from other researchers (as accuracies are the most commonly reported metric). We report our accuracies in Table A.3. We report four types of accuracy depending on the number of accurately predicted dichotomies in each model. To ensure our accuracies are not misleading, we report the results for a model where we sample personality types randomly and a model where we simply select the majority class. We report the theoretical accuracies for the random classifier and we determine the accuracies for the majority classifier using the proportion of types in Figure 3.1. Note that the best models for each metric are bolded.

	Accurately Predicted Dichotomies			
Model	4	≥ 3	≥ 2	≥ 1
Standard	19.69	57.96	88.69	98.67
Weighted	19.63	58.07	**88.93**	**98.80**
Upsampled	**19.70**	**58.16**	88.48	98.76
SMOTE	18.03	55.11	86.12	98.37
Downsampled	11.23	43.75	79.53	96.77
SMOTE + Downsample	18.29	56.02	86.98	98.32
Random Classifier	6.250	31.25	68.75	93.75
Majority Class	15.31	54.54	87.20	98.28

Table A.3: Reported accuracies for the random forests models. For each model, we include the accuracy of correctly predicting all four dichotomies, at least three dichotomies, at least two dichotomies and at least one dichotomy. We report the theoretical accuracies of a random classifier and we determine the accuracies of a majority class classifier using the proportion of types observed in Figure 3.1.

Bibliography

[1] Alam, F., Stepanov, E. A. and Riccardi, G. [2013], 'Personality Traits Recognition on Social Network - Facebook', *Proceedings of the International AAAI Conference on Web and Social Media* **7**(2), 6–9.

[2] Allport, G. W. and Odbert, H. S. [1936], 'Trait-Names: A Psycho-Lexical Study', *Psychological Monographs* **47**(1), i–171.

[3] Altinok, D. [2021], *Mastering Spacy: An End-to-End Practical Guide to Implementing NLP Applications Using the Python Ecosystem*, Packt Publishing Ltd.

[4] Amato, C. H. and Amato, L. H. [2005], 'Enhancing Student Team Effectiveness: Application of Myers-Briggs Personality Assessment in Business Courses', *Journal of Marketing Education* **27**(1), 41–51.

[5] Ambridge, B. [2014], *Psy-Q: You Know Your IQ - Now Test Your Psychological Intelligence*, Profile Books.

[6] Amirhosseini, M. H. and Kazemian, H. [2020], 'Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®', *Multimodal Technologies and Interaction* **4**(1), 9.

[7] Australian Psychological Society [1966], 'Australian Psychologist', *Australian psychologist* .

[8] Aviles, C. B. [2001], *A Review of the Myers-Briggs Type Inventory: A Potential Training Tool for Human Services Organizations*, Education Resources Information Center.

[9] Bagrow, J. P., Liu, X. and Mitchell, L. [2019], 'Information Flow Reveals Prediction Limits in Online Social Activity', *Nature Human Behaviour* **3**(2), 122–128.

[10] Bailey, R. P., Madigan, D. J., Cope, E. and Nicholls, A. R. [2018], 'The Prevalence of Pseudoscientific Ideas and Neuromyths Among Sports Coaches', *Frontiers in Psychology* **9**, 641.

[11] Barabási, A.-L. and Pósfai, M. [2016], *Network Science*, 1st edn, Cambridge University Press, Cambridge, United Kingdom.

[12] Başaran, S. and Ejimogu, O. H. [2021], 'A Neural Network Approach for Predicting Personality from Facebook Data', *SAGE Open* **11**(3), 21582440211032156.

[13] Beach, S. [n.d.], 'Are Personality Tests Valid?', https://www.plum.io.

[14] Bess, T. L. and Harvey, R. [2002], 'Bimodal Score Distributions and the Myers-Briggs Type Indicator: Fact or Artifact?', *Journal of personality assessment* .

[15] Bharadwaj, S., Sridhar, S., Choudhary, R. and Srinath, R. [2018], Persona Traits Identification Based on Myers-Briggs Type Indicator (MBTI) - a Text Classification Approach, *in* '2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)', pp. 1076–1082.

[16] Bird, S., Klein, E. and Loper, E. [2009], *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, 1st edn, O'Reilly Media, Beijing ; Cambridge Mass.

[17] Block, J. [2010], 'The Five-Factor Framing of Personality and Beyond: Some Ruminations', *Psychological Inquiry* **21**(1), 2–25.

[18] Block, M. [2018], 'How the Myers-Briggs Personality Test Began in a Mother's Living Room Lab', *NPR* .

[19] Booth, J. [n.d.], 'Can Your Myers-Briggs Type Change? This Is What to Know If You Manage to Test Differently After a Few Years', https://www.bustle.com/p/can-your-myers-briggs-type-change-this-is-what-to-know-if-you-manage-to-test-differently-after-a-few-years-7829091.

[20] *Botometer Pro Api Documentation (OSOME)* [n.d.], https://rapidapi.com/OSoMe/api/botometer-pro.

[21] Boyle, G. J., Stankov, L. and Cattell, R. B. [1995], Measurement and Statistical Models in the Study of Personality and Intelligence, *in* 'International Handbook of Personality and Intelligence', Perspectives on Individual Differences, Plenum Press, New York, NY, US, pp. 417–446.

[22] Bronsdon, C. [n.d.], 'What Do Different Twitter Emojis Mean?', https://conorbronsdon.com/blog/what-do-different-twitter-emojis-mean.

[23] Calderon, P. [2018], 'VADER Sentiment Analysis Explained'.

[24] Capraro, R. M. and Capraro, M. M. [2002], 'Myers-Briggs Type Indicator Score Reliability Across: Studies a Meta-Analytic Reliability Generalization Study', *Educational and Psychological Measurement* **62**(4), 590–602.

[25] Carroll, R. [2004], 'Myers-Briggs Type Indicator - the Skeptic's Dictionary - Skepdic.com', http://skepdic.com/myersb.html.

[26] Carskadon, T. G. and Cook, D. D. [1982], 'Research in Psychological Type', **5**, 6.

[27] Cattell, H. E. P. [1996], 'The Original Big Five: A Historical Perspective', *European Review of Applied Psychology / Revue Européenne de Psychologie Appliquée* **46**(1), 5–14.

[28] Champion, M. and Krasnolutska, D. [2022], 'Ukraine's Tv Comedian President Volodymyr Zelenskyy Finds His Role as Wartime Leader', https://www.japantimes.co.jp/news/2022/02/26/world/volodymyr-zelenskyy-wartime-president/.

[29] Chen, E. and Ferrara, E. [2022], 'Tweets in Time of Conflict: A Public Dataset Tracking the Twitter Discourse on the War Between Ukraine and Russia'.

[30] Chen, J., Qiu, L. and Ho, M.-H. R. [2020], 'A Meta-Analysis of Linguistic Markers of Extraversion: Positive Emotion and Social Process Words', *Journal of Research in Personality* **89**, 104035.

[31] Cherry, K. [2021], 'What Are the Big 5 Personality Traits?', https://www.verywellmind.com/the-big-five-personality-dimensions-2795422.

[32] Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F. and Wattenberg, M. [2019], 'Visualizing and Measuring the Geometry of BERT'.

[33] Collins, B. [n.d.], 'After Mueller Report, Twitter Bots Pushed 'Russiagate Hoax' Narrative', https://www.nbcnews.com/tech/tech-news/after-mueller-report-twitter-bots-pushed-russiagate-hoax-narrative-n997441.

[34] Collins, B. and Korecki, N. [2022], 'Twitter Bans Over 100 Accounts That Pushed #IStandWithPutin', https://www.nbcnews.com/tech/internet/twitter-bans-100-accounts-pushed-istandwithputin-rcna18655.

[35] Costa, P. and Mccrae, R. [1992], 'Neo PI-R Professional Manual', *Psychological Assessment Resources* **396**.

[36] Cramér, H. [1946], Mathematical Methods of Statistics, *in* 'Mathematical Methods of Statistics', Princeton Mathematical Series ; 9, Princeton University Press, Princeton.

[37] Crawford, S. L. [1989], 'Extensions to the Cart Algorithm', *International Journal of Man-Machine Studies* **31**(2), 197–217.

[38] Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. and Tesconi, M. [2017], The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race, *in* 'Proceedings of the 26th International Conference on World Wide Web Companion', WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 963–972.

[39] De, N. C., Cindolo, L., Sarchi, L., Iseppi, A., Rizzo, M., Riccardo, B., Minervini, A., Sessa, F., Muto, G., Bove, P., Vittori, M., Bozzini, G., Castellan, P., Mugavero, F., Panfilo, D., Saccani, S., Falsaperla, M., Schips, L., Celia, A., Bada, M., Porreca, A., Pastore, A., Yazan, A. S., Marco, G., Novella, G., Rizzetto, R., Trabacchin, N., Guglielmo, M., Pini, G., Lombardo, R., Rocco, B., Antonelli, A. and Tubaro, A. [2020], 'Using a Machine Learning Algorithm to Predict Prostate Cancer Grade', *Journal of Urology* **203**(Supplement 4), e1236–e1236.

[40] De Vries, R. E. [2020], 'The Main Dimensions of Sport Personality Traits: A Lexical Approach', *Frontiers in Psychology* **11**.

[41] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. [2019], 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'.

[42] Dickey, D. and Fuller, W. [1979], 'Distribution of the Estimators for Autoregressive Time Series With a Unit Root', *JASA. Journal of the American Statistical Association* **74**.

[43] Digman, J. M. [1990], 'Personality Structure: Emergence of the Five-Factor Model', *Annual Review of Psychology* **41**, 417–440.

[44] Dimitriu, A. [n.d.], 'Does Nature or Nurture Determine Your Personality? — Psychology Today Australia', https://www.psychologytoday.com/au/blog/psychiatry-and-sleep/202106/does-nature-or-nurture-determine-your-personality.

[45] Divakar, V. [2019], 'Detecting Bots on Twitter Using Botometer', https://blog.quantinsti.com/detecting-bots-twitter-botometer/.

[46] Dixon, S. [2022], 'Number of Worldwide Social Network Users 2027', https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[47] Donaldson, S. I. and Grant-Vallone, E. J. [2002], 'Understanding Self-Report Bias in Organizational Behavior Research', *Journal of Business and Psychology* **17**(2), 245–260.

[48] Donges, J. [2009], ‘What Your Choice of Words Says About Your Personality’, https://www.scientificamerican.com/article/you-are-what-you-say/.

[49] Doroshenko, L. and Lukito, J. [2021], ‘Trollfare: Russia’s Disinformation Campaign During Military Conflict in Ukraine’, *International Journal of Communication* **15**(0), 28.

[50] Druziuk, Y. [2022], ‘A Citizen-Like Chatbot Allows Ukrainians to Report to the Government When They Spot Russian Troops — Here’s How It Works’, https://www.businessinsider.com/ukraine-military-e-enemy-telegram-app-2022-4.

[51] Eysenck, H. J. [1992], ‘Four Ways Five Factors Are Not Basic’, *Personality and Individual Differences* **13**(6), 667–673.

[52] Eysenck, H. J. [1995], *Genius: The Natural History of Creativity*, 1st edn, Cambridge University Press, Cambridge ; New York.

[53] Flores, A. C., Icoy, R. I., Peña, C. F. and Gorro, K. D. [2018], An Evaluation of Svm and Naive Bayes with SMOTE on Sentiment Analysis Data Set, *in* ‘2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)’, pp. 1–4.

[54] Francis, L. and Jones, S. H. [2000], ‘The Relationship Between the Myers-Briggs Type Indicator and the Eysenck Personality Questionnaire Among Adult Churchgoers’.

[55] Gallwitz, F. and Kreil, M. [2022], ‘Investigating the Validity of Botometer-Based Social Bot Studies’.

[56] Gerber, A. S., Huber, G. A., Doherty, D., Dowling, C. M. and Ha, S. E. [2010], ‘Personality and Political Attitudes: Relationships Across Issue Domains and Political Contexts’, *American Political Science Review* **104**(1), 111–133.

[57] Gholipour, B. [2019], ‘How Accurate Is the Myers-Briggs Personality Test?’, https://www.livescience.com/65513-does-myers-briggs-personality-test-work.html.

[58] Giorgi, S., Ungar, L. and Schwartz, H. A. [2021], Characterizing Social Spambots by Their Human Traits, *in* ‘FINDINGS’.

[59] Gjurković, M. and Šnajder, J. [2018], 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’, Association for Computational Linguistics, New Orleans, Louisiana, USA, pp. 87–97.

[60] Golbeck, J., Robles, C., Edmondson, M. and Turner, K. [2011], 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', pp. 149–156.

[61] Goldberg, L. [n.d.], 'The Structure of Phenotypic Personality Traits. - PsycNet', https://psycnet.apa.org/doiLanding?doi=10.1037%2F0003-066X.48.1.26.

[62] Granger, C. W. J. [1969], 'Investigating Causal Relations by Econometric Models and Cross-Spectral Methods', *Econometrica* **37**(3), 424–438.

[63] Grant, A. [n.d.], 'Goodbye to MBTI, the Fad That Won't Die — Psychology Today', https://www.psychologytoday.com/intl/blog/give-and-take/201309/goodbye-mbti-the-fad-won-t-die.

[64] Gray, R. [n.d.], 'The Importance of Personality Trait Screening for Today's Organizations – Application of the Five Factor Model (FFM)', https://sites.psu.edu/leadership/2017/09/02/the-importance-of-personality-trait-screening-for-todays-organizations-application-of-the-five-factor-model-ffm/.

[65] Gunčar, G., Kukar, M., Notar, M., Brvar, M., Černelč, P., Notar, M. and Notar, M. [2018], 'An Application of Machine Learning to Haematological Diagnosis', *Scientific Reports* **8**(1), 411.

[66] Guntuku, S. C., Giorgi, S. and Ungar, L. [2018], *Current and Future Psychological Health Prediction Using Language and Socio-Demographics of Children for the Clpysch 2018 Shared Task*.

[67] Hicks, M. E. [n.d.], The Relationship Between Personality Type and Marital Satisfaction Using the Myers-Briggs Type Indicator and the Marital Satisfaction Inventory, Ed.D., University of North Texas, United States – Texas.

[68] Hindman, M. [n.d.], 'How Cambridge Analytica's Facebook Targeting Model Really Worked – According to the Person Who Built It', http://theconversation.com/how-cambridge-analyticas-facebook-targeting-model-really-worked-according-to-the-person-who-built-it-94078.

[69] Holland, A. S. and Roisman, G. I. [2008], 'Big Five Personality Traits and Relationship Quality: Self-Reported, Observational, and Physiological Evidence', *Journal of Social and Personal Relationships* **25**(5), 811–829.

[70] Hong, L., Convertino, G. and Chi, E. [2011], 'Language Matters in Twitter: A Large Scale Study', *Proceedings of the International AAAI Conference on Web and Social Media* **5**(1), 518–521.

[71] Howlader, P., Pal, K. K., Cuzzocrea, A. and Kumar, S. D. M. [2018], Predicting Facebook-Users' Personality Based on Status and Linguistic Features Via Flexible Regression Analysis Techniques, *in* 'Proceedings of the 33rd Annual ACM Symposium on Applied Computing', SAC '18, Association for Computing Machinery, New York, NY, USA, pp. 339–345.

[72] *How Technology and Social Media Empower the Introvert* [2015], https://knowledge-leader.colliers.com/editor/how-technology-and-social-media-empower-the-introvert/.

[73] Huber, D., Kaufmann, H. and Steinmann, M. [2017], The Missing Link: The Innovation Gap, *in* D. Huber, H. Kaufmann and M. Steinmann, eds, 'Bridging the Innovation Gap: Blueprint for the Innovative Enterprise', Management for Professionals, Springer International Publishing, Cham, pp. 21–41.

[74] Hutto, C. and Gilbert, E. [2015], *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, The AAAI Press.

[75] Ickes, W. [2009], Chapter 8, The Big Five, *in* 'Strangers in a Strange Lab', Oxford University Press, pp. 121–141.

[76] Jolly, M. [n.d.], '(MBTI) Myers-Briggs Personality Type Dataset', https://www.kaggle.com/datasets/datasnaek/mbti-type.

[77] Jones, S. [n.d.], 'PLATO — Computer-Based Education System — Britannica', https://www.britannica.com/topic/PLATO-education-system.

[78] Judge, T. A. and LePine, J. A. [2007], The Bright and Dark Sides of Personality: Implications for Personnel Selection in Individual and Team Contexts, *in* 'Research Companion to the Dysfunctional Workplace: Management Challenges and Symptoms', New Horizons in Management, Edward Elgar Publishing, Northampton, MA, US, pp. 332–355.

[79] Jung, C. G. [1976], *Collected Works of C.G. Jung, Volume 6: Psychological Types*, 1st edition edn, Princeton University Press, Princeton.

[80] *Jung Personality Test: What Is It?* [n.d.], https://www.wikijob.co.uk/aptitude-tests/test-types/jung-personality-test.

[81] Kandler, C. [2012], 'Nature and Nurture in Personality Development: The Case of Neuroticism and Extraversion', *Current Directions in Psychological Science* **21**(5), 290–296.

[82] Keh, S. S. and Cheng, I.-T. [2019], 'Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-Trained Language Models'.

[83] Keller, T. R. and Klinger, U. [2019], 'Social Bots in Election Campaigns: Theoretical, Empirical, and Methodological Implications', *Political Communication* **36**(1), 171–189.

[84] Kielczewski, B. [2013], 'MyPersonality Database', https://www.psychometrics.cam.ac.uk/productsservices/mypersonality.

[85] Klepper, D. [2022], 'Russian Propaganda 'Outgunned' by Social Media Rebuttals', https://apnews.com/article/russia-ukraine-volodymyr-zelenskyy-kyiv-technology-misinformation-5e884b85f8dbb54d16f5f10d105fe850.

[86] Komarraju, M., Karau, S. J., Schmeck, R. R. and Avdic, A. [2011], 'The Big Five Personality Traits, Learning Styles, and Academic Achievement', *Personality and Individual Differences* **51**(4), 472–477.

[87] Kosinski, M. [2018], 'MyPersonality Project Details in a Nutshell', https://sites.google.com/michalkosinski.com/mypersonality.

[88] Lane, H. W., Maznevski, M. L., Mendenhall, M. E. and McNett, J. [2009], *The Blackwell Handbook of Global Management: A Guide to Managing Complexity*, John Wiley & Sons.

[89] Laurence, P. [2022], 'How Ukraine's 'Ghost of Kyiv' Legendary Pilot Was Born', *BBC News* .

[90] Li, W., Chen, Y., Hu, T. and Luo, J. [2018], 'Mining the Relationship Between Emoji Usage Patterns and Personality'.

[91] Magazine, S. and Eveleth, R. [n.d.], 'The Myers-Briggs Personality Test Is Pretty Much Meaningless', https://www.smithsonianmag.com/smart-news/the-myers-briggs-personality-test-is-pretty-much-meaningless-9359770/.

[92] Mathews, P., Mitchell, L., Nguyen, G. and Bean, N. [2017], The Nature and Origin of Heavy Tails in Retweet Activity, *in* 'Proceedings of the 26th International Conference on World Wide Web Companion', WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 1493–1498.

[93] Matthews, B. [2015], 'Best Time to Tweet for Clicks, Retweets and Engagement', https://empower.agency/best-time-to-tweet-clicks-retweets-engagement/.

[94] McAdams, D. P. [1995], 'What Do We Know When We Know a Person?', *Journal of Personality* **63**(3), 365–396.

[95] McCrae, R. R. and Costa, P. T. [1989], 'Reinterpreting the Myers-Briggs Type Indicator from the Perspective of the Five-Factor Model of Personality', *Journal of Personality* **57**(1), 17–40.

[96] Meaker, M. [n.d.], 'This Student's Side Hustle Will Help Decide Musk Vs. Twitter', *Wired UK* .

[97] *Multi-Construct IPIP Inventories* [n.d.], https://ipip.ori.org/newMultipleconstructs.htm.

[98] Muscat, S. and Siebert, Z. [2022], 'Laptop Generals and Bot Armies: The Digital Front of Russia's Ukraine War — Heinrich Böll Stiftung — Brussels Office - European Union', https://eu.boell.org/en/2022/03/01/laptop-generals-and-bot-armies-digital-front-russias-ukraine-war.

[99] Musek, J. [2007], 'A General Factor of Personality: Evidence for the Big One in the Five-Factor Model', *Journal of Research in Personality* **41**(6), 1213–1233.

[100] Myers, I. B., McCaulley, M. H., Quenk, N. L. and Hammer, A. L. [1998], *MBTI Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator, 3rd Edition*, 3rd edition edn, Consulting Psychologists Press, Palo Alto, Calif.

[101] Myers, I. B. and Myers, P. B. [1995], *Gifts Differing: Understanding Personality Type*, 2nd ed. edition edn, CPP, Palo Alto, Calif.

[102] *Myers-Briggs Personality Testing: Understanding How We Relate to the World* [n.d.], http://www.mindtools.com/pages/article/newCDV_51.htm.

[103] *Myers-Briggs Type Indicator® (MBTI®) — Official Myers Briggs Personality Test* [n.d.], https://www.themyersbriggs.com/en-US/Products-and-Services/Myers-Briggs.

[104] Newman, M. [2018], *Networks: Second Edition*, second edn, Oxford University Press UK, Oxford, United Kingdom ; New York, NY, United States of America.

[105] Nguyen, D. Q., Vu, T. and Tuan Nguyen, A. [2020], BERTweet: A Pre-Trained Language Model for English Tweets, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', Association for Computational Linguistics, Online, pp. 9–14.

[106] Nguyen, K. [2022], 'How Putin's Propaganda Is Sowing Seeds of Doubt to Deny Sympathy for Ukraine', *ABC News* .

[107] Nowack, K. [1996], 'Organizational Performance Dimensions', http://www.opd.net/abstracts5.html.

[108] Olan, F., Jayawickrama, U., Arakpogun, E. O., Suklan, J. and Liu, S. [2022], 'Fake News on Social Media: The Impact on Society', *Information Systems Frontiers* .

[109] Orabi, M., Mouheb, D., Al Aghbari, Z. and Kamel, I. [2020], 'Detection of Bots in Social Media: A Systematic Review', *Information Processing & Management* **57**(4), 102250.

[110] Osborne, C. [2022], 'Ukraine Destroys Five Bot Farms That Were Spreading 'Panic' Among Citizens', https://www.zdnet.com/article/ukraine-takes-out-five-bot-farms-spreading-panic-among-citizens/.

[111] Patil, S. M., Singh, R., Patil, P. and Pathare, N. [2021], Personality Prediction Using Digital Footprints, *in* '2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)', pp. 1736–1742.

[112] Paunonen, S. V. and Jackson, D. N. [2000], 'What Is Beyond the Big Five? Plenty!', *Journal of Personality* **68**(5), 821–835.

[113] Pennebaker, J., Boyd, R., Jordan, K. and Blackburn, K. [2015], *The Development and Psychometric Properties of LIWC2015.*

[114] Pennebaker, J. W. and Francis, M. E. [1996], 'Cognitive, Emotional, and Language Processes in Disclosure', *Cognition and Emotion* **10**(6), 601–626.

[115] Peterson, E. L. [2013], 'Myers-Briggs Type Indicator (MBTI) Explained', https://ericlukepeterson.wordpress.com/2013/02/12/myers-briggs-type-indicator-mbti-explained/.

[116] Pittenger, D. [1993], 'Measuring the MBTI ... and Coming up Short', *Journal of Career Planning and Employment* **54**.

[117] Plank, B. and Hovy, D. [2015], 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', Association for Computational Linguistics, Lisboa, Portugal, pp. 92–98.

[118] Platt, J. [2000], 'Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods', *Adv. Large Margin Classif.* **10**.

[119] Polyzos, E. [2022], 'Escalating Tension and the War in Ukraine: Evidence Using Impulse Response Functions on Economic Indicators and Twitter Sentiment'.

[120] Pond, T., Magsarjav, S., South, T., Mitchell, L. and Bagrow, J. P. [2020], 'Complex Contagion Features Without Social Reinforcement in a Model of Social Information Flow', *Entropy* **22**(3), 265.

[121] Pozzana, I. and Ferrara, E. [2020], 'Measuring Bot and Human Behavioral Dynamics', *Frontiers in Physics* **8**.

[122] *Psigometrika by Smit, Gj: Very Good Hardcover (1991) 1st Edition. — Chapter 1* [n.d.], https://www.abebooks.com/first-edition/Psigometrika-Smit-GJ-Haum/3286382544/bd.

[123] *Psychological Testing: Myers-Briggs Type Indicator* [2010], https://www.mentalhelp.net/psychological-testing/myers-briggs-type-indicator/.

[124] Purtill, J. [2022], 'When It Comes to Spreading Disinformation Online, Russia Has a Massive Bot Army on Its Side', *ABC News* .

[125] Quercia, D., Kosinski, M., Stillwell, D. and Crowcroft, J. [2011], Our Twitter Profiles, Our Selves: Predicting Personality with Twitter, *in* '2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing', pp. 180–185.

[126] Rauchfleisch, A. and Kaiser, J. [2020], 'The False Positive Problem of Automatic Bot Detection in Social Science Research', *PLoS ONE* **15**(10), e0241045.

[127] Reed, J. [2007], 'Better Binomial Confidence Intervals', *Journal of Modern Applied Statistical Methods* **6**(1).

[128] Renert, H. [2016], 'Free Myers Briggs Tests Versus the Official MBTI®'.

[129] Robinson, M. [1998], 'How Rare Is Your Personality Type?', https://www.careerplanner.com/MB2/TypeInPopulation.cfm.

[130] Rogers, A., Kovaleva, O. and Rumshisky, A. [2021], 'A Primer in Bertology: What We Know About How BERT Works', *Transactions of the Association for Computational Linguistics* **8**, 842–866.

[131] Rothmann, S. and Coetzer, E. P. [2003], 'The Big Five Personality Dimensions and Job Performance', *SA Journal of Industrial Psychology* **29**(1).

[132] Saifudin, A., Spits Warnars, H. L. H., Soewito, B., Gaol, F., Abdurachman, E. and Heryadi, Y. [2019], *Tackling Imbalanced Class on Cross-Project Defect Prediction Using Ensemble SMOTE*, Vol. 662.

[133] Satow, L. [2021], 'Reliability and Validity of the Enhanced Big Five Personality Test (B5T)'.

[134] Satterfield, K. [2012], 'Use Just a Few Words to Say a Whole Lot'.

[135] Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A. and Menczer, F. [2020], Detection of Novel Social Bots by Ensembles of Specialized Classifiers, *in* 'Proceedings of the 29th ACM International Conference on Information & Knowledge Management', pp. 2725–2732.

[136] Schaubhut, N., Weber, A. and Thompson, R. [2012], 'Myers-Briggs Type and Social Media Report', https://shop.themyersbriggs.com/contents/MBTI_and_Social_Media_Report.aspx.

[137] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P. and Ungar, L. H. [2013], 'Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach', *PLOS ONE* **8**(9), e73791.

[138] Shane, S. [2017], 'The Fake Americans Russia Created to Influence the Election - the New York Times', https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html.

[139] Smart, B., Watt, J., Benedetti, S., Mitchell, L. and Roughan, M. [2022], #IStandWithPutin Versus #IStandWithUkraine: The Interaction of Bots and Humans in Discussion of the Russia/Ukraine War, *in* F. Hopfgartner, K. Jaidka, P. Mayr, J. Jose and J. Breitsohl, eds, 'Social Informatics', Vol. 13618 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, pp. 34–53.

[140] South, T., Smart, B., Roughan, M. and Mitchell, L. [2022], 'Information Flow Estimation: A Study of News on Twitter', *Online Social Networks and Media* **31**, 100231.

[141] Stella, M., Ferrara, E. and De Domenico, M. [2018], 'Bots Increase Exposure to Negative and Inflammatory Content in Online Social Systems', *Proceedings of the National Academy of Sciences* **115**(49), 12435–12440.

[142] Sumpter, D. [2018], *Outnumbered: From Facebook and Google to Fake News and Filter-Bubbles – the Algorithms That Control Our Lives*, illustrated edition edn, Bloomsbury Sigma, London.

[143] Tadesse, M. M., Lin, H., Xu, B. and Yang, L. [2018], 'Personality Predictions Based on User Behavior on the Facebook Social Media Platform', *IEEE Access* **6**, 61959–61969.

[144] Tandera, T., Hendro, Suhartono, D., Wongso, R. and Prasetio, Y. L. [2017], 'Personality Prediction System from Facebook Users', *Procedia Computer Science* **116**, 604–611.

[145] *The Myers & Briggs Foundation - MBTI® Basics* [2022], https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/.

[146] *The Myers & Briggs Foundation - Take the MBTI® Instrument* [2022], https://www.myersbriggs.org/my-mbti-personality-type/take-the-mbti-instrument/.

[147] *The Story of Isabel Briggs Myers - CAPT.Org* [n.d.], https://www.capt.org/mbti-assessment/isabel-myers.htm.

[148] Thomas, T. [2004], 'Russia's Reflexive Control Theory and the Military', *The Journal of Slavic Military Studies* **17**(2), 237–256.

[149] Thompson, E. R. [2008], 'Development and Validation of an International English Big-Five Mini-Markers', *Personality and Individual Differences* **45**(6), 542–548.

[150] Tupes, E. C. and Christal, R. E. [1992], 'Recurrent Personality Factors Based on Trait Ratings', *Journal of Personality* **60**(2), 225–251.

[151] Varvel, T., Adams, S. G., Pridie, S. J. and Ruiz Ulloa, B. C. [2004], 'Team Effectiveness and Individual Myers-Briggs Personality Dimensions', *Journal of Management in Engineering* **20**(4), 141–146.

[152] Vedel, A. [2014], 'The Big Five and Tertiary Academic Performance: A Systematic Review and Meta-Analysis', *Personality and Individual Differences* **71**, 66–76.

[153] Viljoen, S. [n.d.], 'The Validity of the Jung Personality Questionnaire with Reference to Tradesmen'.

[154] Vinney, C. [n.d.], 'Understanding the Big Five Personality Traits', https://www.thoughtco.com/big-five-personality-traits-4176097.

[155] Wakefield, J. [2022], 'Ukraine Invasion: How the War Is Being Waged Online', *BBC News* .

[156] Walsh, B. W. and Holland, J. L. [1992], A Theory of Personality Types and Work Environments, *in* 'Person–Environment Psychology: Models and Perspectives', Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, pp. 35–69.

[157] Wapner, J. [2008], 'He Counts Your Words (Even Those Pronouns)', *The New York Times* .

[158] Watt, J. and Smart, B. [2022], 'Tweets Discussing the Russia/Ukraine War'.

[159] *What Makes a Personality Test Reliable and Valid* [n.d.], https://www.hrprofilingsolutions.com.au/blogs/aus-blog/what-makes-a-personality-test-reliable-and-valid.

[160] Wojcik, S., Messing, S., Smith, A., Rainie, L. and Hitlin, P. [2018], 'Bots in the Twittersphere'.

[161] Wong, E. [2022], 'U.S. Fights Bioweapons Disinformation Pushed by Russia and China', *The New York Times* .

[162] Wylie, C. [2020], *Mindf*ck: Inside Cambridge Analytica's Plot to Break the World*, main edition edn, Profile Trade, London.

[163] Yang, K.-C., Ferrara, E. and Menczer, F. [2022], 'Botometer 101: Social Bot Practicum for Computational Social Scientists'.