Master's thesis

Lars-Magnus Underhaug

From Traits to Threats

Identification of Personality Traits for Individuals at Risk of Radicalisation on Social Media

Master's thesis in Computer Science
Supervisor: Björn Gambäck
February 2022

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Lars-Magnus Underhaug

From Traits to Threats

Identification of Personality Traits for Individuals at Risk of Radicalisation on Social Media

Master's thesis in Computer Science
Supervisor: Björn Gambäck
February 2022

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Social media have become a playground for radical communities, serving as a tool for spreading propaganda and recruiting new individuals to their cause. Vulnerable individuals and curious minds are being exposed to content which they may not have been deliberately perusing, which in turn contributes to their radicalisation process. Despite continuous efforts from social media platforms to restrict the publication of radical content, radical communities are still very much active on social media. Multiple studies have investigated the presence of radical communities on social media and successfully identified radicals on Twitter using methods of machine learning and natural language processing. However, identifying already radicalised individuals may not be the ideal starting point in countering these communities. This Thesis argues that a more natural approach would be to identify individuals before they are drawn into these communities. Understanding the process and factors leading up to radicalisation can provide important insight into the minds of those at risk of radicalisation and hinder further recruitment to radical communities.

By using linguistic cues and methods of machine learning, this Thesis explores the personality traits of individuals considered to be at risk of Islamist radicalisation on Twitter. The personality traits were measured according to the Big 5 personality model for the traits *agreeableness*, *extraversion*, *conscientiousness*, *neuroticism*, and *openness*. A total of three pre-existing personality datasets, consisting of tweets and essays labelled with Big 5 personality scores, were accumulated and used for training personality prediction models. The best performing models were selected and used for personality prediction on a manually collected and annotated dataset consisting of 15,195 tweets from 259 Twitter users believed to be at risk of radicalisation. As a counterpoise, the predicted traits were compared to the traits of ordinary Twitter users, derived from on a randomly sampled dataset of consisting of 25,624 tweets from 259 non-radical Twitter users.

This Thesis contributes by proposing a method for identifying users believed to be at risk of radicalisation on social media, by utilising the social media networks of already radicalised individuals and a set of indicators derived from related work on radicalisation. In addition, this Thesis provides a new to the field, in-depth analysis of the personality traits of Twitter users at risk of radicalisation and how they may differ from ordinary users. The results show that the proposed data collection and annotation scheme is able to successfully identify individuals at risk of radicalisation, yielding an inter-annotator agreement, measured by Cohen's Kappa, of 0.83. The analysis of the predicted personality traits shows that users at risk of radicalisation have common profiles for agreeableness and conscientiousness. When comparing the predicted traits to that of ordinary, non-radical Twitter users, the predictions show a marginal difference in distribution for agreeableness, openness, and conscientiousness, indicating a certain difference in personality between the two domains.

Sammendrag

Sosiale medier har blitt et viktig verktøy for radikale grupperinger som muliggjør spredning av propaganda og rekruttering av individer til radikale miljøer. Gjennom spredningen av slikt innhold vil sårbare og nysgjerrige individer kunne eksponeres for informasjon som de i utgangspunktet ikke har oppsøkt. For enkelte vil denne eksponeringen kunne manifestere seg og lede til såkalt selvradikalisering. Mange sosiale medieplattformer har tatt grep i et forsøk på å begrense spredningen av radikalt innhold. Til tross for disse tiltakene er det, på det tidspunkt denne oppgaven er skrevet, fortsatt flere aktive radikale å finne på sosiale medier. Tidligere studier har vist at radikale brukere kan identifiseres på Twitter gjennom bruk av maskinlæringsmetoder og språkbehandlingsteknikker. Dette er ikke nødvendigvis den mest effektive tilnærmingen for å begrense veksten av radikale miljøer. Denne masteroppgaven argumenterer for at en mer naturlig tilnærming vil være å identifisere potensielle fremtidige radikale forut for radikalisering. Gjennom tilegning av kunnskap om disse individene vil man kunne jobbe proaktivt og begrense rekruttering.

Denne masteroppgaven utforsker personlighetstrekkene til personer ansett å være sårbare for radikalisering i sosiale medier gjennom automatisk predikasjon av personlighetstrekk. Personlighetstrekkene ble målt med utgangspunkt i Big 5-modellen for personlighetstrekkene åpenhet, ekstraversjon, planmessighet, medmenneskelighet, og nevrotisisme . Tre datasett bestående av tweets og essayer merket med verdier for Big 5-personlighetstrekk ble samlet inn og brukt til å trene et utvalg av maskinlæringsmodeller. Videre ble det samlet inn, prosessert og annotert et datasett bestående av 15.195 tweets fra 259 Twitter-brukere ansett å være sårbare for radikalisering. Som en motvekt til dette datasettet ble også det samlet inn et datasett bestående av 25.624 tweets fra 259 ordinære, ikke-sårbare Twitter-brukere. De beste modellene for personlighetspredikasjon ble så valgt ut og brukt til å predikere personlighetstrekk for de to domenene.

Et av hovedbidragene til denne oppgaven er en metode for identifisering av sosiale medier-brukere som anses å være sårbare for radikalisering. Metoden baserer seg på de sosiale nettverkene til allerede radikaliserte brukere og et sett med indikatorer på radikalisering, utformet med utgangspunkt i tidligere forskning. Utover dette bidrar oppgaven med en nyvinnende og grundig analyse av personlighetstrekkene til individer som anses å være sårbare for radikalisering, samt en analyse av hvordan disse personlighetstrekkene skiller seg fra ikke-sårbare personer. Resultatene viser at man gjennom bruk av den fremsatte datainnsamlings- og annoteringsmetoden kan identifisere brukere som er sårbare for radikalisering. Ved sammenligning av annotering fra to uavhengige personer oppnår metoden en innbyrdes annoteringsenighet på 0.83, målt ved Cohens Kappa. Videre analyse av de predikerte personlighetstrekkene viser at personer sårbare for radikalisering har lignende profil for personlighetstrekkene medmenneskelighet og planmessighet. Ved sammenligning av de predikerte personlighetstrekkene mot ikke-sårbare personer viser analysen forskjeller i distribusjon for personlighetstrekkene medmenneskelighet, åpenhet, og planmessighet. Disse resultatene indikerer at personligheten til personer sårbare for radikalisering kan være forskjellig fra ikke-sårbare personer.

Preface

This Master's Thesis written during the fall of 2021, as part of my Master's of Science (MSc) degree in Computer Science at the Department of Computer Science (IDI) at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway.

I would like to thank my supervisor Björn Gambäck for providing interesting articles, thoughts, feedback and for encouraging me to follow up on proposed topics, even when in doubt. I would also like to thank him for his support and understanding during a year out of the ordinary, both in terms of covid and personal challenges. I would also like to thank my girlfriend Frida Julianne Johansen for being there with me through the challenging time of isolation, which without her would have been a much lonelier endeavour. Finally, I would like to thank Jay Alammar, Jacob Devlin, and Maeda-Gutiérrez et al. for permitting the usage and adaption of figures from their papers.

Lars-Magnus Underhaug
Trondheim, 16th February 2022

Contents

List of Figures

List of Tables

1. Introduction

Radicalisation is a gradual social process in which an individual adopts radical or extremist views on political, social or religious issues. Traditionally, radicalisation occurred through physical interactions within social environments, but has later migrated to social media platforms and forums. Individuals associated with radical groups use these platforms to spread hate, violence and abusive content to a global audience, with the aim of recruiting new individuals to their cause. This Master's Thesis focuses on Twitter users found to be vulnerable towards radicalisation and the Big 5 personality traits of these users. A dataset of users is built from a collection of Twitter users and models for personality prediction are built using a set of machine learning algorithms. The predicted personality trait of the users is further analysed to detect similarities among the users with regards to ordinary Twitter users.

The introductory chapter starts off by presenting the background and motivation behind this Master's Thesis along with an explanation of social media services discussed in this Thesis. The following section presents the goal of the Thesis, which is further concretised into four different research questions. Finally, the conducted research method is explained, along with the contributions and structure of the Thesis.

1.1. Background and Motivation

Social media platforms have grown to become an important part of our daily lives, serving as a tool for communicating and expressing thoughts, opinions and news, freely to the public. In 2021, Facebook reported over 2.89 billion monthly active users[1], all from different cultures, backgrounds and parts of the world. Similarly, Twitter reported more than 330 million monthly active users in 2019[2], generating around 500 million tweets per day. Unfortunately, some people take advantage of this freedom of expression by spreading hate, violence and abusive content to a global audience.

Already in 2006, Rogers and Neumann, on commission by the *European Commission's Directorate General for Justice, Freedom and Security*, studied mobilisation factors driving radicalisation within radical Islamist communities in Europe. Their study highlighted the increase of Internet as a tool for recruitment and dissemination of radical propaganda. In 2016, social media played a primary or secondary role in

[1] https://www.statista.com/statistics/241552/share-of-global-population-using-facebook-by-region/

[2] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

radicalising 86.75% of US far-left, far-right, Islamist, and single-issue extremists from PIRUS, a dataset of US extremists [3], compared to 48% in 2012 and 2.86% in 2006 (Jensen et al., 2018). The highest rates were displayed for Islamist extremists with social media playing a role in 93.18% of the radicalisation cases in 2016. Similar metrics are found in the UK, where a study of convicted extremists shows that social media and the Internet played a role in 83% of the studied radicalisation cases in the period 2015-17, 64% in 2010-14, and 35% in 2005-09 (Kenyon et al., 2021). According to a 2015 report by the US Government[4], more than 25,000 individuals left their country to become foreign fighters in Syria. The report also found many of these individuals to be active on social media, both during and after their involvement with extremist organisations, enabling them to stay radicalised even after departing conflict zones.

Despite continuous efforts by these platforms to shut down radical accounts, users may still be exposed to content which they may not have been deliberately pursuing. Identifying radicalisation and already radicalised individuals is a top priority for both counter-extremist agencies and the social media platforms themselves. Several studies have been carried out on detecting extremists online and identifying traits of radicalised individuals. However, most of these previous studies have been concerned with identifying signs of extremism on social media for radicalised users. This might not be the ideal starting point in the fight against extremism as these users may have already made up their minds. Identifying characteristics of people prior to being radicalised can prove useful in hindering the recruitment of new individuals to extremist organisations.

Among many radicalisation researchers, there is believed to be no ultimate profile or single indicator for people likely to adopt radicalised behaviour, and the pathway to radicalisation may unfold differently for every individual[5](Borum, 2011b). Still, social scientists and psychologists have come a long way in developing models for capturing the main factors driving people to radicalisation. One way of utilising these models is by exploring how radicalisation may unfold online. The presence of digital footprints on social media platforms can reveal more about a user than they might realise. Crawling these platforms for footprints has the potential to reveal information about users' personality traits, habits and behaviour.

Models of natural language processing have proven useful for identifying characteristics of individuals. The Big 5 personality model is well known within psychology and has, as evident by Section 6.2.2, been used in several studies on automatic personality prediction, with promising results. Predicting and analysing the personality traits of individuals at risk of radicalisation has the potential to reveal new information about these individuals and help better understand how certain individuals are driven to

[3] https://www.start.umd.edu/data-tools/profiles-individual-radicalization-united-states-pirus

[4] https://www.govinfo.gov/content/pkg/CPRT-114HPRT97200/pdf/CPRT-114HPRT97200.pdf

[5] https://www.southtyneside.gov.uk/article/35878/Young-people-and-radicalisation-and-extremism

extremism. Targeting individuals prior to radicalisation is key to preventing further recruitment to terrorist organisations. Adding knowledge to the factors that may serve as indicators of radicalisation is essential in the fight against extremism.

1.2. Goals and Research Questions

Based on the background and motivation presented in the preceding section, the following goal was formalised for this Master's Thesis:

Goal *Investigate the personality traits traits of individuals at risk of radicalisation on social media by training an automatic personality prediction model using linguistic cues.*

Several models from social science aim to capture the underlying factors contributing to individuals being radicalised, yet there is a lack of research on the personality traits of those being radicalised. By training a model for automatic personality prediction using multiple datasets containing unique posts made by users, annotated with their corresponding Big 5 personality traits, the model will be used for predicting personality traits of individuals at risk of radicalisation on social media. The Thesis will focus on jihadist-oriented radicalisation for Twitter users. In order to reach this goal and guide the research, the goal is concretised into four different research questions. The research questions are presented below.

Research Question 1 *What are indicators of individuals being radicalised on social media?*

The focus of this research question is to explore existing models and indicators of radicalisation based on previous research and social science models. The aim is to identify how radicalisation may unfold in a social media context.

Research Question 2 *How can indicators of radicalisation be modelled for building datasets of users vulnerable to radicalisation on social media?*

The focus of this research question is to explore the field of annotation and use the insight from Research Question 1 to create an annotation scheme for identifying social media users at risk of radicalisation.

Research Question 3 *Do people at risk of radicalisation share any common personality traits and do these traits differ from regular users?*

The focus of this research question is to explore the predicted personality traits of people at risk of radicalisation, to identify similarities and differences among these users. In addition, these users will be compared to a set of users considered not to be at risk of radicalisation to identify any anomalies.

Research Question 4 *Do any indicators of radicalisation correlate more with certain personality traits than others?*

The focus of this research question is to identify the predictive power of the indicators found as part of Research Question 1 and whether these indicators correlate more with certain personality traits than others.

1.3. Research Method

In order to answer the research questions and achieve the goal of this Thesis, a combination of methodologies was applied. For addressing Research Question 1, a qualitative research method was applied. Related work and previous research on radicalisation were explored in the form of a structured literature review, utilising the approach suggested by Kofod-Pedersen (2018). Several recognised models depicting radicalisation were interpreted, along with previous research on computational detection of radicalisation on social media. From this, a set of indicators for radicalisation was derived, suitable for targeting individuals at risk of radicalisation on social media.

Research Question 2 was addressed using a qualitative approach in combination with the knowledge gathered from the preceding research question. Previous research on annotation schemes for radicalisation in social media was explored and a conceptual understanding of the Twitter data was established. On the basis of this knowledge, a data gathering plan was set up and a new annotation scheme was formed. The dataset of users at risk of radicalisation was gathered on the basis of the recommendations made by Parekh et al. (2018), building on the network and mentions from Twitter users identified as being radicalised. Following radicalised individuals on social media is identified as being an indicator of ongoing radicalisation[6] and thus offering a starting point for identifying and creating a dataset of Twitter users at risk of being radicalised. The Twitter stream of the identified users was pulled using Twitter's Academic Research API and each tweet addressed according to an annotation scheme. The tweets were automatically annotated based on a set of indicators for radicalisation, with each indicator being represented by dictionaries containing common terminology for the specific indicator. Tweets containing indicators were then manually verified, with the final annotation being done on a user level. A presentation of reviewed literature and the proposed annotation scheme can be found in Chapter 6 and Chapter 8.

To answer Research Question 3, a combination of qualitative research methodologies and experiments was conducted. In order to find a suitable model for the experiments, research on state-of-the-art models used for personality prediction was reviewed. The model of choice was then trained on multiple datasets annotated with scores for the Big 5 personality traits. A set of Twitter users at risk of radicalisation was then collected and

[6] `https://www.dni.gov/index.php/nctc-newsroom/nctc-resources/item/1945-homegrown-violent-extremist-mobilization-indicators-2019`

annotated using the approach described in Chapter 8. In addition, a set of regular users not exhibiting signals of radicalisation was gathered. The resulting datasets of users were passed to the personality prediction model to derive a set of personality traits for each user. The resulting traits of each user were then analysed with respect to similarities.

The fourth and last research question was answered by evaluating each of the predicted personality traits from the experiment. The indicators from the annotation scheme were included as features for the personality prediction model. Each trait was then compared against each indicator, one at a time, to detect any correlation between the traits and indicators. In addition, the traits were compared against the remaining linguistic features not part of the indicators, to detect other possible indicators of radicalisation.

1.4. Contributions

1. *A thorough overview of radicalisation models and indicators of radicalisation*
2. *A proposed set of criteria for detecting and annotating users at risk of radicalisation on Twitter*
3. *An annotated dataset of users at risk of radicalisation on Twitter*
4. *A thorough analysis of personality traits for people at risk of radicalisation on Twitter and how they may differ from ordinary Twitter users*
5. *An evaluation of the predictive power of the defined radicalisation indicators with regards to the predicted traits, conducted as a correlation analysis*

1.5. Thesis Structure

The remaining Thesis is structured as follows:

Chapter 2 presents relevant background theory needed for understanding the process radicalisation and the theory behind the personality models used in this Thesis and related work.

Chapter 3 presents relevant background theory needed for understanding the technologies used in this Thesis and related work.

Chapter 4 presents models for text representation and concepts related to data annotation used in this Thesis and related work.

Chapter 5 presents some common pre-existing datasets used within research on automatic personality prediction and radicalisation studies.

Chapter 6 elaborates on existing research related to automatic personality prediction and detection of radicalisation on social media.

Chapter 7 explains the data collection and annotation scheme applied for building a dataset of Twitter users believed to be at risk of radicalisation. In addition, the chapter presents the datasets chosen for training models for personality prediction. The chapter concludes by presenting the pre-prosessing steps applied to the manually collected dataset and the personality datasets used.

Chapter 8 explains the architecture of the models used in the experiments of this Thesis.

Chapter 9 presents the conducted experiments, including the experimental plan, setup and results.

Chapter 10 evaluates the research process of the Thesis and discusses the experimental results in light of the proposed research questions and goals. In addition, the chapter presents ethical considerations for future work.

Chapter 11 provides a conclusion on the Thesis, the contributions provided to the field of research, along with propositions for future work.

2. Radicalisation and Personality

This chapter gives an overview of social media platforms and personality models discussed as part of this Thesis, and a presentation of how radicalisation may be depicted on social media. The majority of the sections included in this chapter were written as part of a preliminary study on radicalisation and personality prediction.

2.1. Social Networking Services

Social networking services enable users to interact with each other, build networks and share content on established online platforms. There is a range of platforms that vary in both their format and number of features. Two of the largest platforms today for sharing textual information are Facebook and Twitter. Both of these platforms will be described below.

2.1.1. Twitter

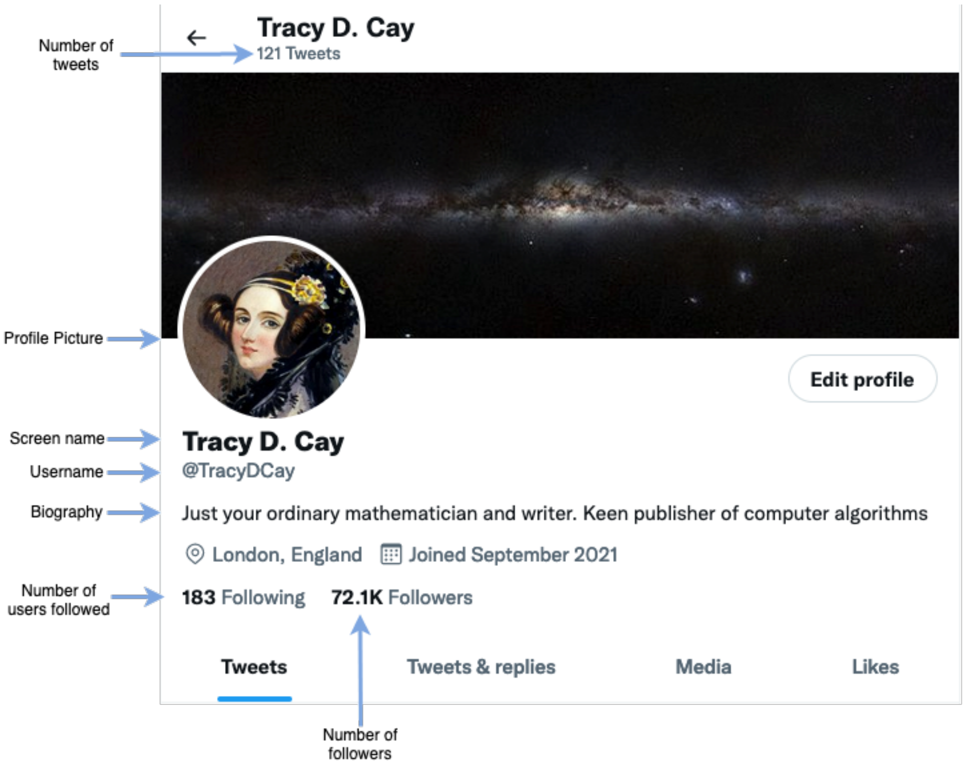
Twitter is a microblogging platform that allows registered users to post, like, or share messages, known as *tweets*. Since its launch in 2006, Twitter has grown to become one of the largest social networking platforms in the world, with more than 330 million monthly active users (as of the 1st quarter of 2019) [1]. Unregistered users are free to use the platform but can only view the content of registered users. The term microblog originates from the term blog, which is a way of conveying information in an informal way via the web. Twitter can thus be viewed as a compressed version of a blog. Twitter is known for its limited and strict format on tweets, which are limited to maximum of 280 characters (increased from 140 characters in 2017). This requires the users to be concise and on point when posting content on the platform.

Twitter allows its users to *follow* other users. This keeps the user updated on all posts made by the followed users. Similarly, a user can be *followed* by other users. The counts of following and followers are displayed on the profile of each user and can be thought of as a virtual social network.

When registering on the platform each user must provide their own unique username. The user is then free to upload a profile picture, header picture, or write a short biography about themselves. This information, in addition to follower count, following

[1] `https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/` (Accessed 11.10.2021)

count and tweets, is displayed on the users own profile. Figure 2.1 shows an example of how such a profile might look like.



Tracy D. Cay
121 Tweets

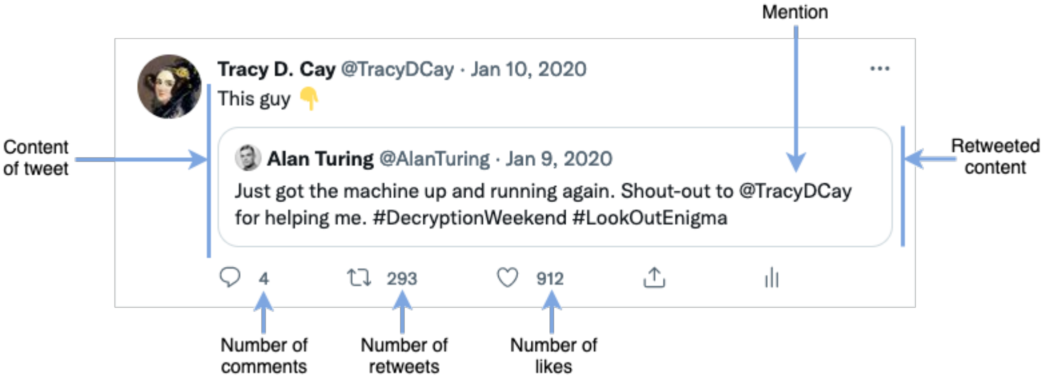
Tracy D. Cay
@TracyDCay
Just your ordinary mathematician and writer. Keen publisher of computer algorithms
London, England Joined September 2021
183 Following 72.1K Followers

Figure 2.1.: Example of a Twitter Profile

Tweets can contain a variety of content, such as images, videos, texts, URLs, hashtags, and mentions. Hashtags are identified by the # symbol and are used to put the tweet within a context, category or highlight a given topic. Users can then access tweets of a specific topic by searching for a given hashtag. This makes finding communities or debate topics on Twitter easy.

Mentions are identified by the @ symbol followed by the username of the person you want to mention. Mentions are used to tag another registered user in the tweet or target the tweet towards a specific user within the Twitter network.

Another common feature on Twitter is *retweeting*. A retweet is a way of reposting another user's tweet, thus adding it to your own tweet stream. When retweeting, the user is free to attach their own tweet to the retweet. Retweeting is one of the ways commonly used to communicate on Twitter, in addition to tweeting and leaving comments on tweets. Figure 2.2 shows what a tweet may look like.



Tracy D. Cay @TracyDCay · Jan 10, 2020

This guy 👇

Alan Turing @AlanTuring · Jan 9, 2020

Just got the machine up and running again. Shout-out to @TracyDCay for helping me. #DecryptionWeekend #LookOutEnigma

Figure 2.2.: Example of a Tweet

2.1.2. Facebook

Facebook is one of the most popular social networking services in the world and the largest by its sheer number of users. The platform was launched in 2004 and as of 31st December 2020, Facebook reported more than 2.8 billion monthly active users [2]. Facebook lets users register their own profile with a personal timeline. The users are free to customise their profile with information such as profile pictures, biographies, demographic information, occupation, interests and hobbies. Users can also upload a variety of content, ranging from images, texts and videos, which can be shared within their own network, consisting of *friends*, or to the public. Initially, Facebook was designed as a platform for sharing content within each user's individual network. Today, Facebook has been expanded with a public news feed where users can view content from numerous sources based on their *followings*, such as groups, organisations, news outlets or even targeted advertisement. The platform has generated users from all over the world with North America ranking highest with 68.5% of the population registered as active Facebook users. The global outreach of the platform is around 28.5%. [3]

2.2. Radicalisation

Understanding the processes that govern radicalisation is crucial for identifying potential future radicals on social media. Radicalisation often starts with open conversations on social media platforms, before moving onto private messages or forums with targeted individuals (Fernandez et al., 2018). Modelling radicalisation is a field of substantial research within social science, psychology, computer technology, and policing. There are numerous models which aim to capture the underlying factors that contribute to people being radicalised. It is however important to emphasise that most of the theories written

[2] `https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx` (Accessed 14.02.2022)

[3] `https://www.statista.com/statistics/241552/share-of-global-population-using-facebook-by-region/` (Accessed 14.02.2022)

about radicalisation into violent extremism may be viewed as conceptual theories. Few of these models have been subject to any scientific or empirical verification. Some of the most established and recognised conceptual models and empirical research will be described below. The selected models are based on observable factors on social media. The reviewed models and research are influenced by the workings of Borum (2011a).

2.2.1. Conceptual Models of Radicalisation

In 2003, Borum proposed the *Process of Ideological Development* model, sometimes referred to as *Borum's Four-Stage Model of the Terrorist Mindset* for identifying factors that may lead to radicalisation. The stages of the model were based on an analysis of violent extremist groups, across multiple ideologies. The model describes radicalisation as a process consisting of four consecutive stages that ultimately lead to an individual being radicalised; radicalisation starts off with a person identifying an event or condition that is considered unfair or not right, referred to as *grievance*. Examples of such events could be poverty, government-imposed restriction, or unemployment. People at this stage display a propensity towards being radicalised, which leads them on to the second step, *comparison* or *injustice*. Here, the identified event is framed as unfair in comparison to others, leading to feelings of inequality. The third stage involves *attribution*, where the inequality is blamed on a nation, policy or person. People reaching this stage are considered to be in the process of indoctrination. The last step, *reaction*, where the target believed to be at fault of the inequality is demonised and dehumanised to justify aggresion towards the target. People reaching this stage falls under the category extremists.

In 2005, Moghaddam proposed a radicalisation model which bared a close resemblance to that of (Borum, 2003). The model was referred to as *The Staircase to Terrorism* and depicts six steps progressively leading an individual to being radicalised. The first three steps of the model are more or less an exact image of the *Process of Ideological Development*, where radicalisation starts off with *feelings of deprivation*, which the person tries to alleviate itself from. This, in turn, leads to *perceived options to unfair treatment* and *displacement of aggression*, where the person becomes frustrated with the current situation and place the blame on a target believed to be at fault for this frustration. The consecutive steps are *moral engagement* where the person may come to sympathise with radicals acting against the target, *legitimacy of the terror organisation* where the person may join a radical movement engaging in terrorism, and *the terrorist act*, where the the sympathiser joining the radical movement commit terrorist acts. The higher up the staircase, the fewer people ascend to the successive level, leaving a small number of people who actually commit terrorism. People are considered to be at risk up until the fourth level, which represents the indoctrination, similar to the third stage of Borum (2003).

The Staircase to Terrorism has been criticised by Lygre et al. (2011) as being to narrow-minded by claiming radicalisation as a linear path. Lygre et al. (2011) suggests that the pathway to radicalisation may have several trajectories and does not follow one single path. Despite the criticism, some of the most circulated models among law enforcement groups assume this linear path. Both Precht (2007) and Silber and Bhatt (2015) describe radicalisation, or the pathway to jihadisation, as a four-stage linear process; Radicalisation starts off with the *pre-radicalisation* phase were the individual is frustrated with their life, society or government. The individual then moves on to identify itself with radical Islam, leading to *conversion*. The third stage involves *indoctrination* in the radical ideology, leading to the final stage which is *jihad*. Though these models are highly conceptual and abstract, they have similarities with Moghaddam and Borum, namely the feeling of frustration or grievance.

In their study of the different roots of radicalisation, Fernandez et al. (2018) reviewed several of the models presented above. In an effort to draw consensus from the models, they bridged radicalisation as a singular and several paths by looking at the driving forces leading to radicalisation and not the path itself. They derived a model consisting of three factors, or roots, namely; *micro* or self-affect root, which is the feeling of deprivation, injustice or threats that motivate an individual to seek out radical movements, *meso* or community root, where the individual meets like-minded people and forms relationships within a community, and *macro* which represents the influence which a society or government has upon a nation or community. If the influence threatens the identity of the community, the latter root can feed to a us-versus-them thinking and in turn fuel the radical community.

2.2.2. Empirical Research on Radicalisation

The amount of empirical studies on radicalisation is sparse. Most of the studies conducted are qualitative in nature and often based on small samples of individuals. This section draws on the most important findings from recent empirical research on radicalisation that translates to social media. In 2006, Slootman et al. explored the early stages of radicalisation among Muslims in Amsterdam. They surveyed 24 young Muslims at the verge of radicalisation, in addition to 12 already radicalised Muslims. Their study revealed two common factors as drivers for radicalisation; 1) an orthodox religious stand with Islam, and 2) mistrust in the current order, with feelings of discrimination towards the Muslim community. In their study of radicalisation in the US and UK, Gartenstein-Ross et al. did an empirical examination of 117 homegrown jihadists. Among their findings, and similar to that of Slootman et al., were the feeling of discrimination towards Muslims. The feeling was represented by a conspiracy that Western society deliberately tries to subjugate Islam and that the West and Islam therefore is incompatible.

In 2010, Rezaei and Goli conducted a 108-item survey of 1,113 Danish immigrants in an effort of explore factors of radicalisation. Among their findings, they found the most radical individuals to be more dissatisfied with life and lonelier than other people.

While these results were interesting, they also support the notion produced by the conceptual models, presented in Section 2.2.1, that the feeling of frustration or grievance to be prevalent among those undergoing radicalisation. Their findings also showed that these individuals were more likely to have experienced discrimination for being Muslim, offering support to the findings of Slootman et al. (2006) and Gartenstein-Ross et al. (2009).

Among other factors found to contribute to radicalisation is the use of social media and the Internet Rogers and Neumann (2007). It has been found to appeal to curious minds and facilitate the radicalisation of lone-wolfs. Social media as a key contributor to radicalisation has been taken seriously by several law enforcement agencies and in 2019 The National Counterterrorism Center (NCTC), in collaboration with FBI and the Department of Homeland Security published a set of indicators of mobilisation in to violent extremism[4] published by *The National Counterterrorism Center*. The indicators were derived from both empirical and conceptual theory, based on peer-reviewed academic studies and brain-storming. Among the indicators were; 1) Consuming or sharing violent extremist videos, media, and/or messaging, linking and taking part in sharing extremist propaganda, 2) Promoting violent extremist narratives, and 3) Communicating directly with violent extremists online.

Upon analysing radicalisation risk on social media Lara-Cabrera et al. (2017) explored the prevalence of indicators associated with radicalisation for the *How ISIS Uses Twitter* dataset, later described in Section 5.1. They found the indicators *feelings of discrimination*, *negativity towards the West*, and *pro-jihadism* rhetoric to be a common theme among the users. They also found the indicators *frustration* and *use of negative words or hate speech* to be prevalent, but with a higher degree of variance among the users.

2.3. Modelling Personality

Modelling personality traits of individuals is a field of substantial research within psychology. There are numerous models which aim to capture the underlying factors contributing to one's personality. Some of the most established and recognised models will be described below.

2.3.1. The Big 5 Personality Model

The Big 5 personality model, sometimes referred to as the Five-Factor model, is a popular model for personality prediction within psychology. Research on modelling personality according to the The Big 5 was first carried out by Fiske (1949) and consisted of four

[4] `https://www.dni.gov/files/NCTC/documents/news_documents/NCTC-FBI-DHS-HVE-Mobilization-Indicators-Booklet-2019.pdf`

factors. The model was later expanded and refined by Goldberg (1990) and McCrae and John (1992). The model structures the personality traits of individuals in a hierarchical manner and consist of five traits; Openness, Extraversion, Conscientiousness, Neuroticism, and Agreeableness. The traits live on a continuous scale between two extremes, where the individual is assigned scores for each trait based on standardised questionnaires. The model is said to capture all human personality traits, across observers and cultures.

- **Openness** describes the level of openness to new experiences and ideas. Individuals scoring high on this trait tend to be curious and have many interests, while individuals scoring low are more resistant to change and tend to resist new ideas.
- **Extraversion** describes the level of sociability and assertiveness. Individuals scoring high on this trait tend to like socialising and making new friends, while individuals scoring low often prefer solitude and dislike small-talk.
- **Conscientiousness** describes the level of planned behaviour and impulsiveness. Individuals scoring high on this trait often spend more time on preparation and are more detail oriented, while individuals scoring low tend to be more impulsive and procrastinate more.
- **Neuroticism** describes the level of emotional stability. Individuals scoring high on this trait are often more confident and calmer, while individuals scoring low tend to be more nervous and anxious.
- **Agreeableness** describes the level of prosocial behaviour. Individuals scoring high on this trait are often friendlier and emphatic towards others, while individuals scoring low tend to be more suspicious and hostile. describes the level of prosocial behaviour. Individuals scoring high on this trait are often friendlier and emphatic towards others, while individuals scoring low tend to be more suspicious and hostile.

The Big 5 personality traits of individuals are captured and quantified by the use of standardised questionnaires. There are several versions of these questionnaires, both long and short. The Big Five Inventory is a questionnaire consisting of 44 statements where individuals can self-report their personality traits. NEO PI-R is another personality test used to assess personality traits of individuals according to The Big 5 Model. The International Personality Item Pool (IPIP)[5] also offers a questionnaire for The Big 5 Model.

2.3.2. The Dark Triad of Personality

The Dark Triad of Personality comprises personality traits related to malevolent human behaviour (McHoskey et al., 1998). As indicated by the name, the model consists of three traits, namely *Psychopathy*, *Machiavellianism* and *Narcissism*. Psychopathy depicts signs of impulsiveness, selfishness, lack of empathy, and anti-social behaviour. Machiavellianism

[5] `https://ipip.ori.org/`

is characterised by manipulation, cynicism, high levels of self-interest and the lack of morality. Narcissism is characterised by lack of empathy, egocentric behaviour, unfounded pride, and grandiosity.

2.3.3. Myers-Briggs Type Indicator

The Myers-Briggs Type Indicator (MBTI) is a psychological self-reporting personality assessment Myers (1998). The test consists of four scales represented by letters; Extroversion (E) or Introversion (I), Judging (J) or Perceiving (P), Sensing (S) or Intuition (I), and Thinking (T) or Feeling (F). A letter is assigned for each scale, yielding a total of 16 possible personality type combinations. An individual characterised as ISTP, according to the model, would be considered Introvert, Sensing, Thinking and Perceiving.

3. Machine Learning for Text Classification

3.1. Machine Learning

Machine learning is a subfield of artificial intelligence that concerns generalisation of data. The goal of a machine learning model is to automatically learn patterns in data to predict future unseen data instances, while continuously self-improving the model (Mitchell, 1997). Machine learning is divided into three categories; supervised learning, unsupervised learning and reinforcement learning (Russell and Norvig, 2009). Supervised learning learns a function for mapping input data to an output label, based on pre-labelled training instances. Supervised learning can be either a classification problem or a regression problem. Classification problems map input to a discrete set of pre-defined classes, while regression problems map the input to continuous values. Unsupervised learning work with unlabelled data instances. The aim of the unsupervised model is to learn structures of the data and find hidden patterns. Reinforcement learning learns by having a so-called actor making actions in a defined environment or action space. The aim of the actor is to reach a goal state or move closer to it. Actions that bring the actor closer to the goal state are rewarded, while actions that brings it further from the goal state are penalised. The task of personality prediction and radicalisation prediction are instances of supervised machine learning, and the most common algorithms will be described below. Most of the sections contained in this chapter were written as part of the preliminary study for this Master's Thesis.

3.1.1. Linear Regression

Linear regression is one of the simplest machine learning algorithms within supervised learning. Linear regression uses a statistical method to solve regression problems by approximating a linear function from numerical independent input variables (Russell and Norvig, 2009). As illustrated by Figure 3.1, linear regression works by fitting a line to the data points by establishing the relationships between independent and dependent variables. This is done by minimising the error between the fitted line and the data points. When single independent variables are used, the prediction is referred to as Simple Linear Regression. When there are two or more independent variables, the prediction is referred to as Multiple Linear Regression.

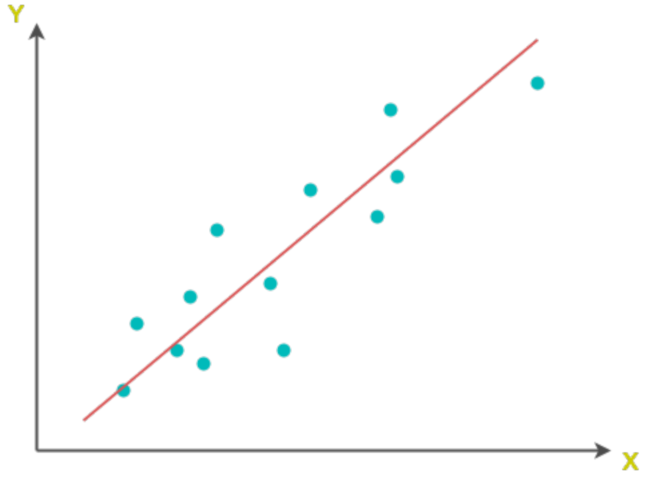


Figure 3.1.: Visualisation of a Linear Function Approximated by Linear Regression

There are several implementations of linear regression for approximating the linear function with minimal error. The *Least Squares* approximation works by minimising the squared error of the training data. The best fitted line is found by minimising the sum of squares of the difference between the observed data point and the fitted value. *Ridge Regression* is a regression algorithm well suited for multiple linear regression where the independent variables are highly correlated (multicollinearity). *Least Absolute Shrinkage and Selection Operator* (LASSO) is a regression algorithm that performs both variable selection and regularisation. LASSO is similar to Ridge Regression but differs by coefficients being approximated to zero more frequently.

3.1.2. Logistic Regression

Logistic regression in one of the more popular supervised machine learning algorithms. The algorithm can be used to solve both regression and classification problems but is mainly used as a classification algorithm where it predicts the probability of an instance belonging to a class(Russell and Norvig, 2009). Logistic regression uses independent variables to predict categorically dependent variables (discrete variables). The discrete variables are predicted by using a logistic function or a sigmoid function, as illustrated by Figure 3.2.

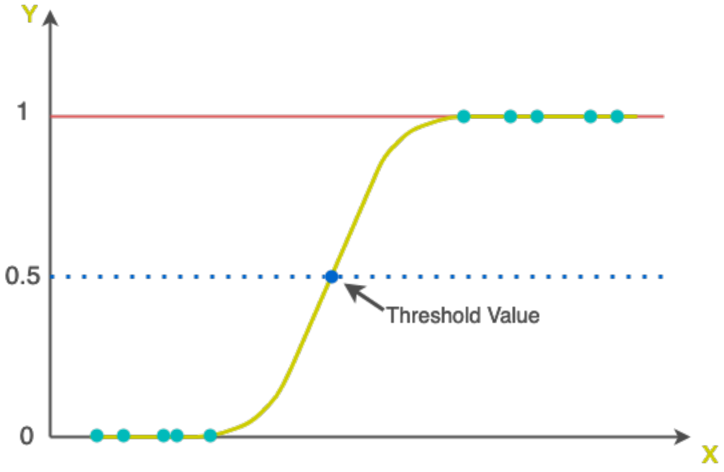


Figure 3.2.: Visualisation of Function Approximated by Logistic Regression

The function maps every real number to 0 or 1, based on a threshold value, indicating which class the instance belongs to. The predictive function is based on Equation 3.1, where x represents the probability of an instance belonging to a class. For machine learning, a parameterised version of logistic regression is used.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

3.1.3. Support Vector Machines

The Support Vector Machine is a supervised learning model applicable in both classification and regression problems (Russell and Norvig, 2009). Here, the classification implementation of the model will be explained. The SVM is fed a set of training instances labelled with corresponding targets. The SVM works by fitting a hyperplane to the data that separates it into distinct classes, while maximising the margin to each class's training instances. An example of SVM is given in Figure 3.3. Each training instance is represented with coordinates in n-dimensional vector space, where n is given by the number of features and the dimensions of the hyperplane is given by *n-1*. The data points closest to the hyperplane are referred to as *support vectors*. The distance between the support vectors and the hyperplane is referred to as the *margin*. The SVM aims to separate the classes by maximising the margin between the two support vectors. The resulting hyperplane is used for classifying new instances to their respective classes. If the data points are not linearly separable by the hyperplane, they are mapped to a higher dimension using a kernel function.

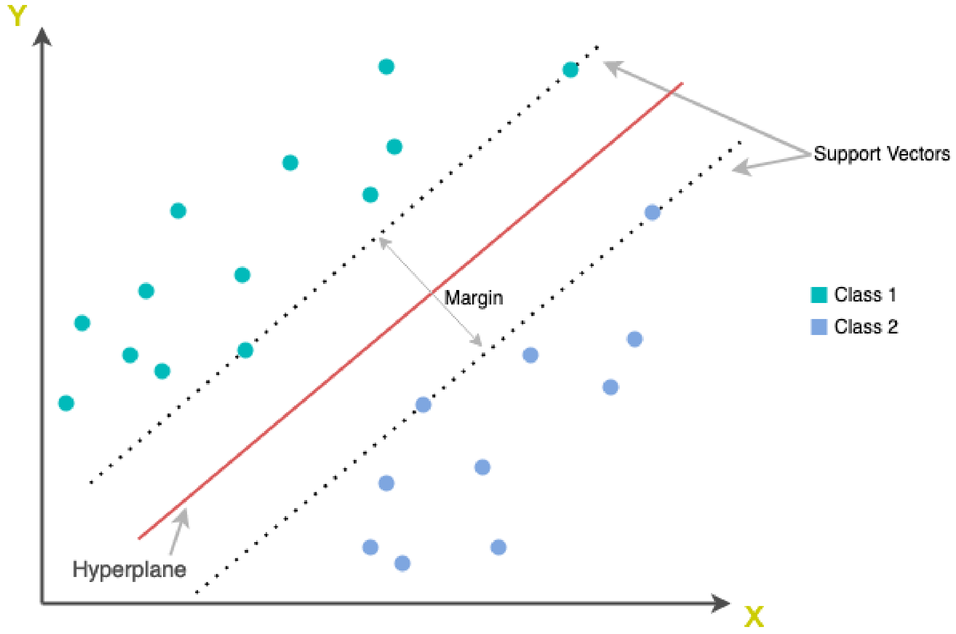


Figure 3.3.: Visualisation of Two Classes Separated by SVM Hyperplane

3.1.4. Naïve Bayes Classifier

The Naïve Bayes classifier is based on Bayes theorem (Bayes and Price, 1763) and assumes conditional independence between the training instances. Naïve Bayes works by computing the probability of an instance belonging to a certain class. Decision rules are applied and the instance is assigned to the class with the highest probability. The assumption of conditional independence is not always met, but the classifier is still able to perform reasonably well on these instances.

3.1.5. Decision Trees

Decision trees is a classification model that builds a tree-like structure for assigning class labels to data instances (Mitchell, 1997). Features are represented as nodes in the tree. Edges from nodes represent all possible choices that can be taken from that node, given as the value of each feature. The range of possible classes is represented as leaf nodes in the tree. New instances are classified by traversing the tree, from root to leaf, based on the feature values of the instance. The class corresponding to the ending leaf node is then assigned to the instance. Decision trees are often used in combination, where the collection of decision trees is referred to as the Random Forest algorithm. New instances are classified by taking the predictions of each individual decision tree and combining them to obtain a final class label. Random forest is often chosen over decision trees, as building a single best decision tree can prove to be challenging.

3.1.6. Gaussian Processes

Gaussian Processes (GP) is a supervised kernel-based machine learning algorithm for solving both regression and classification tasks. As opposed to some machine learning algorithms which typically outputs a predicted value or class, GP takes a probabilistic approach and produces a probability distribution for the output values by fitting a kernel function (Rasmussen and Williams, 2005). This makes GP well suited for problems such as regression tasks and short text classification in combination with word embeddings (Ma et al., 2015). The idea behind GP is based on similarity, measured by the kernel. If if two input values or vectors appear close together in the input space, they are more likely to be close together in the output space. The likelihood is represented as a probability distribution, or a Gaussian distribution, given as $p(f(x_1), f(x_2, ..., f(x_n)$. GP is dependent on a mean function and kernel function, or covariance function, in order to produce a final output value for $f(x)$. Both functions are given by Equation 3.2 and Equation 3.3.

Kernel function (covariance):

$$K(x_1,\ x_2) = \mathbb{E}[(f(x_1) - m(x_1)) \times (f(x_2) - m(x_1))] \tag{3.2}$$

Mean function:

$$m(x) = \mathbb{E}[f(x)] \tag{3.3}$$

3.1.7. Kernels

Multiple linear machine learning algorithms, such as Support Vector Machines and Gaussian Process, depend upon a kernel function, some times referred to as the *kernel trick*, in order to solve non-linear problems. A kernel function is a function which maps data points onto a higher dimension. By doing so, the linear classifier is able to separate non-linear data. The kernel function works by using the dimensional input, represented as the feature vectors of the space, and outputting the dot product of the data in the space. Linear classifiers can be implemented with different kernels and the choice of kernel often depend on the task at hand. Some of the more popular kernels are the *Radial Basis Function (RBF)*, *simoid*, and *polynomial* kernel. Sigmoid and polynomial similar and both calculated using the product of two feature vectors, one being transposed, while RBF is calculated using the euclidean distance between two feature vectors. The kernel functions, represented as K, are given by Equation 3.4-3.6 and are based on the official Scikit Learn documentation[1].

Simoid kernel:

$$K(x,y) = \tanh(\gamma x^T y + c_0) \tag{3.4}$$

[1] `https://scikit-learn.org/stable/modules/metrics.html#sigmoid-kernel`

Polynomial kernel:

$$K(x,y) = \tanh(\gamma x^T y + c_0)^d \tag{3.5}$$

RBF kernel:

$$K(x,y) = \exp(-\gamma||x - y||^2) \tag{3.6}$$

- x and y are feature vectors
- γ is known as a slope
- c_0 is known as the intercept
- d is the kernel degree

3.1.8. Deep Learning

Deep Learning (DL) is a subfield of machine learning based on Artificial Neural Networks. The growth in available data and rise in computational power over the recent years have spiked the interest in DL. DL has the advantage of being able to learn high-level features from raw data, without the need for much manual feature extraction. This has proven to be useful when working with domain specific texts, such as social media data. Raw data is fed through a weighted network of perceptrons, to learn representations and find hidden patterns in data automatically. The structure of the perceptrons is based on neural science and is a high-level abstraction of human neural cells, which are activated whenever sufficient stimulation is achieved. The remaining section is dedicated to the following DL concepts; Artificial Neural Networks, Recurrent Neural Networks, Convolutional Neural Networks, Encoder-Decoder Architecture and the Attention mechanism, Transformers and Bidirectional Encoder Representation from Transformers.

Artificial Neural Networks

As explained in Section 3.1.8, Artificial Neural Networks (ANN) form the basis for a lot of deep learning implementations. The structures of ANNs can vary substantially depending on the problem, where the simplest form is a feed-forward network consisting of nodes and weighted edges structured in a layered fashion (Mitchell, 1997), as illustrated in Figure 3.4. In a feed-forward network, the information flows in one direction, from input to output.

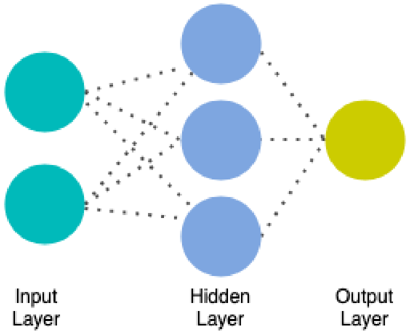


Figure 3.4.: Visualisation of the ANN architecture

ANNs can be constructed with multiple hidden layers or no hidden layers at all. A network with multiple layers is called a Multi-Layer Perceptron (MLP), whereas a network with no hidden layers is called a Single-Layer Perceptron (SLP). The lack of hidden layers in SLPs limits these networks to linearly separable functions. MLPs, however, are able to work with high-dimensional data and thus overcome the limitations of SLPs. The number of hidden layers and nodes are often found through experimentation and fine tuning of the network and are often heavily dependent on the dataset and problem at hand.

ANNs work by each node being connected to its subsequent layer by weighted edges. The propagation of information through each node is based on an activation function. The resulting output, passed as input to the subsequent layer, is normally a product of the propagated information and the weight of the corresponding edge (O'Shea and Nash, 2015). The network is continuously optimised by training on new data instances. A common optimisation technique is *back-propagation*, where the output of the network is compared to the optimal output, before passing the calculated error back through the network to adjust weights and improve the network. The training continues until the network converges.

Recurrent Neural Networks

Recurrent Neural Networks (RNN) is an extension of ANNs, where information is allowed to flow in cycles, as opposed to feed-forward networks. This is enabled by each node having its own internal state for storing information about previous calculations(Mitchell, 1997). The stored information is used to produce output based on both past and current decisions, in which the most recent decisions weigh more heavily. The memories of RNNs are however limited, making them unsuitable for data with long-term dependencies, such as long texts.

Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a regularised version of MLPs, that exploits hierarchical patterns in data (O'Shea and Nash, 2015). This allows the model to learn more complex relationships in the data with smaller and simpler patterns. This is achieved by having the receptive field of neurons overlap partially. Much like ANNs, CNNs consist of an input layer, a series of hidden layers, and an output layer, but the hidden layers consist of a series of convolutional and pooling layers in addition to a fully connected ANN, as illustrated by Figure 3.5. The convolution layer works as a filter on the input data, to produce maps of activation. The filter can then be applied across the input to detect features. This works by putting the convolutional layers in series to produce a feature map which indicates the locations and strengths of the input features. The pooling layers perform dimensionality reduction on the data by combining the outputs of neurons in a layer into one single neuron for the consecutive layer. In addition to dimensionality reduction, the pooling layer performs regularisation to prevent overfitting and enable the layer to capture more local information. This has proven useful for Natural Language Processing tasks, where single words can inherit important semantic meaning.

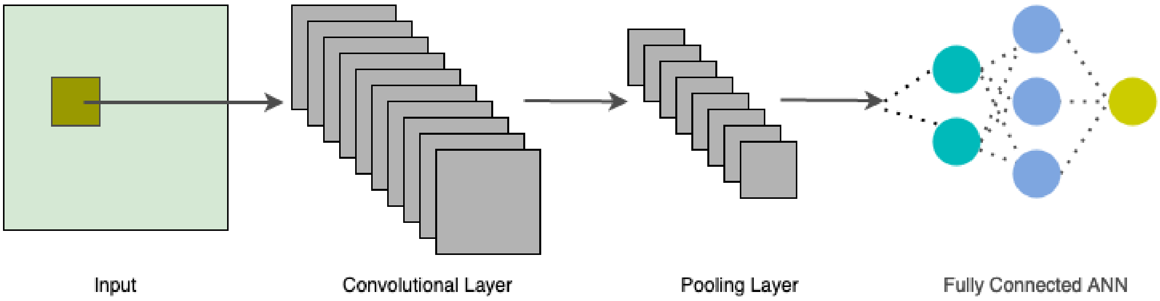


Figure 3.5.: Visualisation of the CNN architecture. Adapted from Maeda-Gutiérrez et al. (2020) Fig. 2, published under CC BY 4.0 License.

The Encoder-Decoder Architecture and Attention Mechanism

The encoder-decoder architecture is an architecture designed for the task of handling sequences of variable length. As given by its name, the architecture consists of two major components, namely the encoder and the decoder. The encoder's task is to take in input sequences of variable length and encode them to a specific fixed shape. The task of the decoder is to map the encoded shape back into a sequence of variable length. The idea is simple but has proven useful in sequence-to-sequence predictions and variations of machine translation. The architecture is often used in combination with other DL models, such as RNNs. However, these combinations have proven to be bad at long-term dependencies, with the performance degrading as the input grows larger.

In an effort to counteract these limitations, Bachrach et al. (2012) proposed a technique for considering the relative importance of words. The technique resulted in the *attention mechanism*, an architecture for quantifying the interdependencies between input and output sequences. The interdependence between input and output sequences is referred to as general attention. The interdependence between the input sequences is referred to as self-attention, which tries to compute a sequence representation from different positions within the sequence. The goal of the attention mechanism is to identify which parts of the input sequence relate to the output. This is achieved by giving the decoder access to previous states of the encoder.

Transformers and Bidirectional Encoder Representations from Transformers

Vaswani et al. (2017) built upon the ideas proposed in Section 3.1.8 to create the Transformer architecture. The Transformer architecture is able to handle large sequential data by combining both the attention mechanism and the encoder-decoder architecture. Unlike previous implementations of the encoder-decoder architecture, the transformer architecture does not depend on recurrence or convolutions, making it parallelisable. This enables transformers to work on larger datasets. Transformers work by having several encoders encode the input, while the attention mechanism produces contextual information on the input by weighing the relevance of the input. The decoder then processes the encoded information using the incorporated contextual information, to produce output.

The high degree of parallelisation of transformers has made them preferable for task involving larger corpora and given rise to new machine learning models. One of these models is Bidirectional Encoder Representations from Transformers (BERT). BERT is a machine learning model developed at Google (Devlin et al., 2019), designed to pre-train bidirectional representations from unlabelled text by conditioning both the right and left context of all layers, utilising the transformer architecture. BERT is pre-trained on a large corpus consisting of 2,500 million words from the English version of Wikipedia and 800 million words from the Toronto Book Corpus. The pre-training is done for Masked Language Modeling, where missing tokens and next sentences are predicted from a set of input tokens and a pair of sentences. The combination of transformers and substantial pre-training makes BERT able to produce state-of-the-art models for a variety of tasks, such as question answering and NLP tasks, without the need for much modification. This makes BERT a very popular transfer learning model.

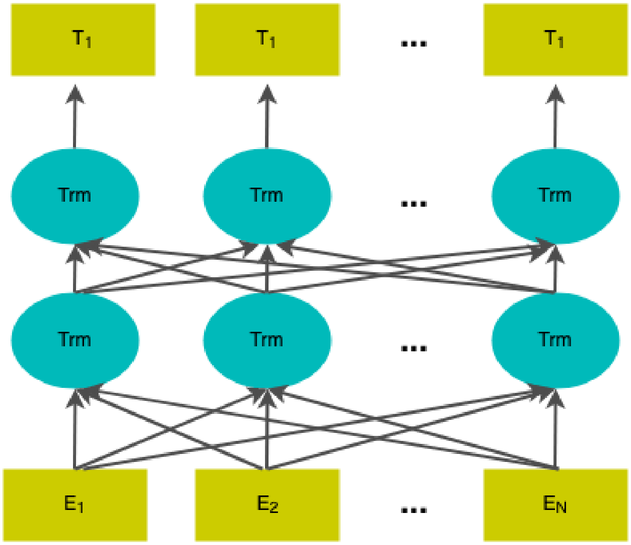


Figure 3.6.: Visualisation of the BERT architecture. Adapted from Devlin et al. (2019), with permission from Jacob Devlin

The basic architecture of BERT is given by Figure 3.6. E_1, $E_2 ... E_N$ is the first layer of the architecture and represents the word embeddings of each term. The word embeddings are passed on to the consecutive *Trm* layer, which performs a multi-headed attention computation to produce new intermediate representations. The intermediate representation can be passed on through several *Trm* layers, before being passed to the final T_1, $T_2 ... T_N$ layer which produces the final contextualised representation of the input. There have been introduced several newer implementations of the BERT architecture, among them Facebook's replication, RoBERTa Liu et al. (2019), and the reduced version, ALBERT (Lan et al., 2020).

3.1.9. ALBERT

While several pre-trained transformer-based models such at GPT-3 and BERT are capable of achieving state-of-the-art performances on several NLP tasks, they often come with a major drawback; their size. Models such as $\text{BERT}_{\text{LARGE}}$ cannot be run on normal computers as they require huge amounts of GPU memory, which in turn limits their adoption possibilities. In an effort to tackle these limitations, Lan et al. (2020) introduced A Light BERT (ALBERT). ALBERT bases its architecture on BERT, but proposes three new improvements: factorised embedding parameterisation, cross-layer parameter sharing, and inter-sentence coherence loss. Factorised embeddings parameterisation decomposes the embedding parameters (of size E) into two matrices, from BERT's O(V x H), to O(V x E + E x H), where V is the size of the vocabulary and H is the size of the hidden layers. This two-stage mapping makes for a substantial parameter reduction when H » E. The second improvement is the cross-layer parameter sharing, meaning both both self-attention and feed-forward parameters are shared across all twelve layers. This results in a substantial decrease in the number of parameters, while keeping up the performance. The third improvement is the removal of BERT's Next Sentence Prediction (NSP), as it had been proven unreliable by the authors. Instead, sentence-order prediction loss is added, which together with topic prediction, enables inter-sentence coherence prediction.

3.1.10. DistilBert

DistilBERT (Sanh et al., 2020), as the name suggest, is a distilled version of BERT. Similar to that of Lan et al. (2020), the motivation behind DistilBERT was the growing demands for computational power that came with the increased size of emerging transformer models. The idea DistilBERT is thus simple; reduce the size of the pre-trained model while maintaining the performance of BERT. DistilBERT is able to retain 97% of the performance of $\text{BERT}_{\text{BASE}}$, while being 60% faster. DistilBERT is able to achieve this performance through two steps; knowledge distillation and transfer learning. Knowledge distillation is essentially a compression technique where the larger model, meaning BERT, is compressed into a smaller model, meaning DistilBERT. This compression results in the number of layers being reduced from 12 to 6 and in turn the number of parameters being decreased by 40%, from the original $\text{BERT}_{\text{BASE}}$ to DistilBERT. Transfer learning occurs by DistilBERT being trained on the same dataset as BERT. BERT's loss function is used to monitor the loss function of DistilBERT, thus making DistilBERT able to reproduce the behaviour of BERT.

3.1.11. Ensemble Learning

Ensemble learning is a technique used within machine learning, rather than a model. Instead of relying purely on one model to make its prediction, ensemble learning uses a combination of weaker base models to produce a final predictive model. The final predictive model then bases its predictions on a combination of all predictions made by the base models Russell and Norvig (2009). The technique can be compared to deciding whether your car needs fixing or not. Instead of asking for the advice of one mechanic, you refer to several mechanics for their opinion. Your final decision may be based on several assessments, but the simplest way would be a majority voting. There are several ways to make the final prediction, e.g. adding a neural network as a final predictor. Some popular variations of Ensemble learners are *Bagging* and *Boosting*. Bagging involves training multiple base models on different subsets of the training data, where the final prediction is given as an average of the predictions. Boosting involves adding the base models sequentially, where a models are able to correct the predictions of a preceding model. The final prediction is given as a weighted average of the predictions.

Ensemble learners may be implemented using similar base models or different base models. The choice of base models may depend upon the selected ensemble learning method. When boosting is implemented, the ensemble model can benefit from diversity between base models. It is assumed that different base models will make different mistakes, thus offering the possibility to correct these mistakes.

3.2. Evaluation Metrics

A crucial part of developing machine learning models is evaluating the performance of the models. There are several metrics for evaluating performance and some of the most common will be described below.

3.2.1. Accuracy

Accuracy is perhaps the simplest and most intuitive metric for evaluating the performance of machine learning models. Performance is measured by quantifying the number of correctly classified instances over the total number of instances. Accuracy is given by Equation 3.7.

$$\text{Accuracy} = \frac{tp + tn}{tp + fp + tn + fn} \tag{3.7}$$

As given Table 3.1, *tp* or true positives is the number of correctly classified instances. *tn* or true negatives is the number of correctly classified false instances. *fp* or false positives is the number of incorrectly classified positive instances. *fn* or false negatives is the number of incorrectly classified negative instances.

Table 3.1.: Confusion Matrix for Binary Classification

		True Value	
		Positive	Negative
Predicted Value	Positive	*tp*	*fp*
	Negative	*fn*	*tn*

3.2.2. Precision and Recall

Precision and *recall* are used to capture the relevance of the classification. Precision is measured by quantifying the proportion of correctly classified positive instances over the total number of positive instances. Precision is given by Equation 3.8.

$$\text{Precision} = \frac{tp}{tp + fp} \tag{3.8}$$

Recall is measured by quantifying the proportion of correctly classified instances over the total number of possible positive instances. Recall is given by Equation 3.9.

$$\text{Recall} = \frac{tp}{tp + fn} \tag{3.9}$$

3.2.3. F-score

The F-score is a metric that combines both precision and recall, and conveys the trade-off between them. The F-score is measured on a scale between 0 and 1, where 1 indicates perfect precision and recall. F-score is given by Equation 3.10.

$$\mathrm{F}_\beta = \frac{\textit{precision} \times \textit{recall}}{(\beta^2 \times \textit{precision}) + \textit{recall}} \times (1 + \beta^2) \tag{3.10}$$

The value of β may vary, but a common value is 1, yielding the F_1-score. The F_1-score, sometimes referred to as the Sørensen-Dice coefficient (Sørensen, 1948; Dice, 1945), is a harmonic mean of precision and recall.

3.2.4. Mean Square Error, Mean Absolute Error, and Root Mean Square Error

The aforementioned evaluation metrics are all related to classification problems. For regression problems, Mean Square Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) can all be used for evaluating the performance of a model. MSE, MAE, and RMSE are used to measure the distance between the predicted value and the expected (or true) value. MSE measures the average square distance between the predicted value and the expected value, while MAE measures the absolute average distance between the predicted value and the expected value. RMSE is the square root of MSE and proves useful as larger errors get penalised more than smaller ones.

3.2.5. Pearson Correlation Coefficient

The Pearson correlation coefficient (Pearson, 1895) measures the linear correlation between two variables. The coefficient is measured on a scale between -1 and 1, where -1 indicates a perfect negative correlation between the two variables, and 1 indicates a perfect positive correlation between the variables. A coefficient value equal to 0 indicates no correlation between the two variables. The Pearson correlation coefficient proves useful for analysing correlations among features and between features and target values.

3.2.6. Area Under the Curve

The Area Under the Curve (AUC) measures a classifier's ability to distinguish between different classes. AUC works by measuring the two-dimensional area underneath the Receiver Operating Characteristic (ROC) curve on a scale between 0 and 1. The higher the AUC, the better the classifier is at distinguishing between positive and negative classes. The ROC curve plots the *True Positive Rate* versus the *False Positive Rate* at different classification thresholds, thus separating signal from noise. AUC then provides an aggregate measure of the performance of the classification model across all classification thresholds.

4. Text Representation and Annotation

4.1. Natural Language Processing

Natural Language Processing (NLP) is a subfield within linguistics, computer science and artificial intelligence that concerns the processing and interpretation of natural language. Natural language is an ever-evolving language that differs from logical language by being ambiguous, redundant and unstructured Russell and Norvig (2009). This makes NLP both a challenging task and a field of substantial research. With social media comes vast amounts of data eligible for NLP, giving rise to new algorithms, feature extraction methods and processing tools. Within text processing, an instance of a text, such as a Facebook status update or a tweet, is referred to as a *Document*. A collection of documents is referred to as a *Corpus*. This section describes some of the methods used for processing natural language texts from social media.

4.1.1. Text Preprocessing

Preprocessing is considered an important preliminary step to prepare a corpus for feature extraction and further analysis, e.g. before training the data on a classifier (Russell and Norvig, 2009). The aim of preprocessing is to remove noise from the corpus, while keeping important information and emphasising features. Balancing between noise and information can be a challenging task and often requires domain specific knowledge. However, there are a few common steps that can be applied to the corpus to better prepare it for analysis.

Segmentation is the task of separating the document text into sentences and is applied prior to any other preprocessing steps. *Tokenisation* can then be applied to divide each sentence into smaller chunks, referred to as tokens. Tokens are sequences of characters that are considered to be meaningful semantic units of the document and are often referred to as terms or words.

Normalisation can be applied to the document and refers to the task of converting words and terms to a common canonical representation. *Stemming* is a common normalisation technique to removes affixes from words, leaving only the stem of the word. An example of this would be the words "drinking" and "drinks". These words would both be reduced to the word "drink" by applying stemming. This illustrates some of the limitations of stemming. Two words that inherit different semantic meaning, one being a verb and the other being a noun, are reduced to the same stem. In these instances, stemming can cause confusion and in turn worsen the performance of the classifier. For this reason, stemming should always be considered carefully. *Lemmatisation* is another common normalisation technique that group words with similar headwords or dictionary form (lemma) together. An example of this would be the word "feet". This word would be interchanged with "foot", from plural to singular. Other normalisation techniques include converting words to lowercase, removing or replacing numbers by their textual representation, and stop words removal. Stop words are words that appear frequently in texts, such as "and", "the", and "for". Stop words contribute little to the overall meaning of a text. Stop words removal can reduce the corpus size and leave meaningful words. However, stop words can be domain specific, and removing them should always be considered carefully. Take for instance the word "not" which is often considered to be a stop word. For semantic analysis, removing "not" from the corpus can change the meaning of a sentence or even a whole text. An example of this could be the sentence "I am not in love with you". Removing "not" from the corpus would change the meaning of this sentence, potentially causing confusion in later analysis.

Social media texts often include unconventional characters, misspellings, elongates words, URLs, and other textual information that can be considered noise. *Noise removal* is a final preprocessing step that can be applied to better prepare the text for further analysis. Noise removal from social media texts can include removing whitespaces, converting URLs, hashtags and mentions to common terms, removing HTML-tags, etc. Noise removal differs from normalisation and tokenisation by being highly domain specific. It is therefore important to know the domain being analysed to reduce the risk of removing important information. For instance, the words "good" and "goooood" can inherit different semantic meaning, with the latter being more expressive than the first, due to the use of elongations. Reducing the latter to its lemma could result in sentiments being lost. An example of a noise removal step could be to reduce the elongated words to a size of three, thus keeping the original sentiment of the word.

4.1.2. Text Representations and Feature Selection

Most machine learning algorithms require the transformation of text to numerical representations in the form of features. These features can be instance counts, lexical- or context-based, metadata, etc. The remaining section will present feature models commonly applied to NLP problems.

Term Frequency-Inverce Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) is a term weighing scheme to measure a terms importance within a corpus (Leskovec et al., 2020). Term Frequency (TF) measures the number of times a term appear within a single document, and is offset by the Inverse Document Frequency, measuring the number of times a term appears across the corpus. Words that appear frequently in the corpus will be punished and of less importance. The weight of terms can be calculated by Equation 4.1 and represented as part of vectors.

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF} * log(\frac{N}{d_f}) \tag{4.1}$$

Bag-of-Words

Bag-of-Words (BoW) gets its name from Harris (1954) and is an approach that encodes text without any consideration of the order or relationship between the words. The approach records the occurrence of words within a document and represents them as a vector of size equal to the vocabulary. Words can be represented based on presence in the document (one-hot-encoding) or Term Frequency, counting the number of times the word occurs within the document. BoW works well in domains where the presence of words represents the content of the document.

n-grams

n-grams is a method, proposed by (Markov, 2006), for representing text as a sequence of characters or words of size *n*. *n*-grams works much like BoW, but the vocabulary is represented by *n*-grams and the order of words are preserved, at least to some degree. *n*-grams works by sliding a window of size *n* over the text, grouping sequences together based on the size of *n*. *n*-grams of size 1 consist of single words or characters (unigrams), size 2 (bigrams) consist of two words or characters, and so forth.

Word Embeddings

Word embeddings is a method based of the distributional hypothesis formalised by Firth (1957) which states; *"You shall know a word by the company it keeps."*. The underlying idea of the hypothesis is that words with similar semantic meaning are distributionally similar. Word embeddings harvest this idea by analysing words that occur in similar settings. Words in the vocabulary are converted to a real valued vector representation where similar words appear closer together in the vector space. There are several popular implementations for constructing word embeddings.

Word2Vec is a word embedding model created at Google and published in 2013 (Mikolov et al., 2013). Word2Vec converts words to a vector representation by one of two proposed methods; Continuous Bag-of-Words (CBOW) or Continuous Skip-Grams. The CBOW architecture learns word embeddings by predicting words based on context. Continuous Skip-Grams learns word embeddings by predicting surrounding words given the current word. The methods differ by continuous skip-grams working better with small datasets and rare words, while CBOW works better with frequent words. Both of the architectures use the local context for creating the vector representation.

Global Vectors and Word Representation (*GloVe*) is an extended version of Word2Vec developed at Stanford University (Pennington et al., 2014). GloVe is an unsupervised algorithm that uses the global context for building a matrix of co-occurring words across the corpus and how frequently they appear together.

fastText is a word embedding model created by Facebook (Bojanowski et al., 2017). The model differs from the previous ones by using single characters as the base for vector representations and representing words as the sum of the character representations. The use of characters for vector representation enables the model to capture morphemes.

Word embeddings can be created based on a corpus, making them domain specific, or one can use word-embeddings pre-trained on a general corpus. The size of the word embeddings will depend upon the dimensions of the vector space. A larger vector space is able to store more information. However, the size of the word embeddings is not proportional to the performance, and a larger vector space is not guaranteed to yield higher performance.

Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) is a language analysis application for categorising words in text into predefined categories (Tausczik and Pennebaker, 2010). LIWC consists of over a range of categories, depending on the version, that words are grouped according to. The application then calculates the percentage of words in each category. The percentage distribution can be used for analysing sentiments or derive semantic meaning from texts. Examples of predefined categories are affect, occupation, pronouns, positive emotions, and negative emotions.

Byte-Pair Encoding

Byte-Pair Encoding (BPE) (Sennrich et al., 2016) is a sub-word segmentation tokenisation algorithm using corpus statistics to segment a text into tokens. The algorithm is used by several machine learning learning models and among them as part of ALBERT's SentencePiece (Kudo and Richardson, 2018) tokeniser. Instead of tokenising the text into words based on whitespaces, the algorithm utilises the data in the text by representing it as a balanced combination of both characters, subwords and words. This makes it suitable for handling both large corpora and out-of-vocabulary words. The algorithm works in in to steps. First, the algorithm creates a corpus representation by pre-tokenising the text into words and count the frequency of each individual words. The second step involves splitting the corpus into single symbols to form what is referred to as the base vocabulary. BPE then works in an iterative manner by merging the symbol pairs that appear most frequently together, extending the base vocabulary to a fixed size.

WordPiece

WordPiece (Schuster and Nakajima, 2012) is a sub-word tokenisation algorithm which bares a close resemblance to that of Byte-Pair Encoding (BPE). The algorithm has gained popularity by being integrated as part of BERT and DistilBert's tokenisation algorithms. WordPiece utilise the data in the text to generate a deterministic representation of characters, subwords and words of the text. This makes it ideal for handling out-of-vocabulary words. The algorithm is given a corpus to train on, along with a parameter specifying the desired number of tokens. The algorithm then works in a similar fashion to that of BPE by splitting the corpus into words and count their frequency. The words are then split into single symbols to form a base vocabulary. The base vocabulary is then extended using a greedy approach. WordPiece iterates over each instance of the base vocabulary making up a pair to calculate their score. WordPiece differs from BPE in the way the score for each candidate token is calculated. Instead of using frequency as a score, the score is calculated as the frequency of the pair, divided by the factor of the frequency for the first and second element of the pair. The score calculation is given by equation 4.2.

$$\frac{\textit{freq of pair}}{\textit{freq of first element} \times \textit{freq of second element}} \tag{4.2}$$

4.2. Annotation

When working with supervised machine learning algorithm, the model requires labelled data that properly conveys the information captured in each data instance. The labels are used for training the model, meaning understanding the relationship between input and output, to predict labels for new unseen data instances. Annotation is the process of analysing and extracting information from a dataset used to label each instance. The annotation can be done automatically by algorithms, manually by humans, or by using a combination of the two, known as semi-supervised learning. The method of choice depends very much upon the task at hand and the information available for the data.

4.2.1. Automatic Annotation

Automatic annotation is a method in which a computer analyses the data, predicts and assign labels to each data instance. The method is often applied when the dataset is to large to be manually analysed or when the data is to complex for humans to understand. In order for the computer to annotate the data, it need some sort of pre-existing knowledge about the labelling of data. This knowledge can be a smaller, already labelled dataset for training a supervised learning model or a set of annotation rules applied by the computer. Regardless of the method, automatic annotation comes with a degree of error and a smaller sample often needs to be analysed in order to verify the correctness of the labels. The accuracy achieved for the automatic annotation model comes down to the similarity between the training data and the data to be annotated.

4.2.2. Manual Annotation

Manual annotation is a method in which a human analyses the data and assigns labels to each data instance accordingly. Manually annotating datasets is both expensive and time-consuming, making it unsuited for larger datasets. More often than not, manual annotation is used for creating smaller samples of data to train supervised learning models for automatic annotation. The extent to which the manual annotation is carried out is dependent on the performance of the trained model. Once the sample is large enough to produce reliable and accurate predictions by the model, the manual annotation usually ends.

The annotation is often based on a number of predefined labels and assigned according to the annotators own interpretation of the data. This however leaves the chance of the annotation being biased or containing mistakes. In an effort to reduce uncertainty, manual annotation is often performed by several annotators. This enables the annotators to verify each others labelling and reach a consensus based on the predefined annotation scheme. In order to reach an inter-annotator agreement, different metrics can be used to measure the similarity, and thus the reliability, of the annotated data. Two common metrics for measuring inter-annotator agreement are Cohen's Kappa and Fleiss' Kappa.

4.2.3. Cohen's Kappa

Cohen's Kappa, referred to as *kappa*, is a metric introduced by Cohen (1960) that measures the inter-agreement between two annotators. The metric differs from percentage calculations for agreement by taking into account the probability of the agreement being a result of chance. This is believed to make the metric more robust and reliable than the regular percentage formula. Cohen's Kappa is calculated using the formula displayed in equation 4.3.

$$\kappa = \frac{p_0 + p_c}{1 - p_c} \tag{4.3}$$

where

$\mathbf{p_0}$ is the relative proportion of agreement between the annotators
$\mathbf{p_c}$ is the proportion of agreement expected by chance

The inter-agreement is measured on a nominal scale with a score ranging between -1 and 1, where 0 indicates agreement by chance and a score of 1 indicates perfect agreement between the annotators. Negative values are unlikely to occur, but indicates that there is no agreement between the annotators. An interpretation of the level of agreement is given by Table 4.1.

Table 4.1.: Interpretation of Cohen's Kappa Landis and Koch (1977)

Kappa Statistic	**Strength of Agreement**
< 0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

Although κ is considered to be a robust measure for inter-agreement, there are drawbacks to the metric that should be taken into account. First, it can only be used for measuring the inter-agreement between two annotators. If a dataset is annotated by more than two annotators, measuring the inter-agreement would call for a different metric. Second, disagreement between annotators are treated the same, making the metric unsuited where there are categories more closely linked to each other than others. Third, in the cases of rare categories, the metric may underestimate the inter-agreement.

4.2.4. Fleiss' Kappa

Fleiss' Kappa was introduced by Fleiss et al. (2003) and is a metric for measuring the inter-agreement for any fixed number of annotators. Fleiss' Kappa is built upon the formula first introduced by Cohen (1960), displayed in equation 4.3, with a few adjustment, making it suitable for multiple annotators. Both both $\mathrm{p_0}$ and $\mathrm{p_c}$ are calculated differently and the equations are presented below.

$\mathbf{p_c}$ is given by the following formula:

$$\mathrm{p_c} = \sum_{j=1}^{k} p_j^2 \tag{4.4}$$

where

k is the total number of labels
j is the index of the label assigned to p_j
$\boldsymbol{p_j}$ is the proportion of all elements that were assigned to the *j*-th element. This is calculated by the following formula:

$$\mathrm{p_j} = \frac{1}{Nn} \sum_{i=1}^{k} n_{ij} \tag{4.5}$$

where

N is the total number of elements to be labelled

n is the number of annotations per element
n_{ij} is the number of *i*-th elements assigned to the *j*-th label.

$\mathbf{p_0}$ is given by the following formula:

$$\mathrm{p_0} = \frac{1}{N}\sum_{i=1}^{k} p_i \tag{4.6}$$

where

$\mathbf{p_i}$ is a measure of the agreement between annotators for the label of the *i*-th element. $\mathrm{p_i}$ is given by the following formula:

$$\mathrm{p_i} = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}^2 - n_{ij} \tag{4.7}$$

5. Pre-Existing Datasets

The task of predicting personality traits of people vulnerable to extremism of social media is dependent on two types of datasets. First, a dataset containing users' streams from social media annotated with personality traits is used to train a model for classification or prediction. Second, in order predict personality traits of people at risk of radicalisation on social media, a dataset containing such users is needed in order for the personality prediction model to output personality traits. Social Media Platform offer vast amounts of data, that when utilised, can offer tremendous insight into the habits of its users. Such data is easily accessible today, either through established datasets or through Application Programming Interfaces (APIs) provided by the platforms themselves. This chapter will give an overview of some pre-existing and popular datasets that have been used in studies on automatic personality prediction and detection of individuals showing signs of radicalised behaviour on social media.

5.1. Radicalisation Datasets

Substantial efforts have been invested into researching ways of using automated systems for detecting extremism and radicalised behaviour on social media. These studies have mainly been concerned with identifying already radicalised individuals based on their social media profiles and activity. Little effort has been invested into identifying people at risk of being radicalised on social media. For this reason, this research field has suffered from a lack of established consensus on how datasets should be labelled to best identify individuals at risk of radicalisation. Of the studies found on the subject, the annotation practice ranges from labelling based of hypotheses to labelling founded in social science. Fernandez et al. (2018) highlighted in their work the need for a gold standard dataset of social media users vulnerable to social media extremism that could be used for training regression and classification models. Such a dataset would need to be manually verified by experts to ensure the quality of the data. Creating such goldstandard datasets may however prove challenging as Social Media Network sites continuously patrol the web, suspending accounts spreading violent propaganda supporting their cause.

Due to limited research on people vulnerable to extremism on social media and the lack of agreement on annotation schemes, care must be taken when using established datasets or labelling manually collected data. This section will give an overview of some of pre-existing datasets with regards to individuals at risk of radicalisation on social media. Some of these datasets are used across multiple studies but processed differently. While these datasets consist primarily of individuals labelled as already radicalised, they have proven useful as a basis for identifying those at risk of radicalisation through network analysis.

5.1.1. O'Callaghan et al. (2014)

The first dataset to be reviewed is that of O'Callaghan et al. (2014). Their study addressed the issue of mapping out different communities within the anti-regime of the Syrian Conflict. Their dataset was comprised of Twitter accounts associated with the Syrian conflict. The accounts were identified through utilising the *Twitter user list* feature, a feature often used by journalists to locate topic experts within the Twitter user base. Their approach focused on identifying users according to two categories:

1. Syrian accounts by known journalists
2. Official accounts of high-profiled entities known to be active on the ground in the Syrian conflict, such as ISIS and the Free Syrian Army (FSA).

O'Callaghan et al. (2014) approach resulted in a total 17 lists, yielding a total 911 unique accounts. The lists were aggregated into a single set and manually analysed, to filter out non-Syrian accounts, and retain accounts claiming to be directly involved in the Syrian conflict. After filtering, the final dataset consisted of 652 accounts, including additional data such as their followers, following, and tweets, all gathered through the Twitter API. In addition, all valid YouTube URLs were extracted from the tweets, and profile data for the videos gathered through the YouTube Data API resulting in a set of 14,629 unique YouTube channels related to 619 of the Twitter accounts.

Accounts	Tweets	Mentions	Retweets
652	1,760,883	175,969	27,768

Table 5.1.: Associated Twitter Data

The dataset was categorised by employing a unified graph approach (cosine similarity and k-Nearest Neighbor) based on information from both Twitter and YouTube. The approach led to the identification of 16 communities, that was later divided into three major categories based on inter-community linkage. These categories were:

1. C-Jihadist: Smallest of the three communities with a total of 40 accounts. Consisting mainly of violent jihadists.
2. C-revolutionary: Consisting of 105 accounts, heavily linked to the Free Syrian Army (FSA).
3. C-moderate: The largest community consisting of 137 accounts, mainly being the moderate portion of the Syrian opposition.

5.1.2. Rowe and Saif (2016)

In their 2016 study, Rowe and Saif aimed to analyse the radicalisation signals of social media users prior to and post radicalisation. In order to reach their goal, they constructed a dataset of Pro-, Anti- and Neutral-ISIS users resided in Europe. To build their dataset they leveraged the dataset of O'Callaghan et al. (2014) consisting of 652 users pertaining the Syrian conflict. These users served as seed accounts to generate a new dataset. Rowe and Saif (2016) started off by checking which of the 652 users were still active and had their timelines accessible on Twitter. The filtering resulted in 512 seed accounts from which their followers was collected. The resulting collection consisted of 2.4 million users.

The dataset was pruned down to users believed to be located in Europe by looking for unique European location names and countries by string matching user's biography with these locations. After filtering, the dataset was left with around 154K users.

As the study was concerned with analysing the behaviour of users prior to them being radicalised, the study would need a complete timeline of each user. At the time, the Twitter API posed a limit on the number tweets that could be retrieved from a user's timeline, to the latest 3,200 tweets. Based on the limit, Rowe and Saif (2016) were able to retrieve the full timeline for 97% of the users. This resulted in a dataset containing around 104M tweets. The study chose to focus on English and Arabic tweets as they made up most of the collection (43% English and 41% Arabic) and the authors were fluent in both languages. The users were labelled as pro-, anti- or neutral-ISIS based on the following hypothesis:

- H1: Sharing of ISIS related material on Twitter
- H2: Use of extremist related language in tweets

Both hypotheses were based on literature reviews of radicalisation research. For H2 the authors manually constructed a dictionary of pro- and anti-ISIS terms in both English and Arabic. The dictionary was based of related work and conferring with religious experts. Users were then labelled as becoming activated (radicalised) once fulfilling H1 or having at least five separate tweets fulfilling H2. After applying the hypothesis to the set of 154K users, the authors were left with the following:

- 508 H1 users
- 208 H2 users
- 727 users within the union of H1 and H2
- 64 users in the intersection between the H1 and H2 sets.

5.1.3. Fifth Tribe - How ISIS uses Twitter

The *How ISIS Uses Twitter*[1] is a dataset created by **Fifth Tribe**, a digital agency serving businesses, non-profit organisations and government agencies. The dataset was created as a reaction to the 2015 Paris attacks and consists of around 17,000 tweets from over 100 pro-ISIS fanboys. The dataset was part of an initiative to develop effective counter-messaging measures against violent extremism all around the world. The dataset was built over a three-month period, by gathering data based on keywords such as Dawla, Wilayat, etc. and filtered based on images and the users Twitter network.

The dataset has proven useful in several studies, among them in workings of Fernandez et al. (2018) and Lara-Cabrera et al. (2017). The dataset consists of the following information for each of the users:

- Name
- Username
- Description
- location
- Number of followers
- Number of tweets
- Timestamps of tweet
- Tweet messages

5.1.4. Tweets Targeting ISIS

The *Tweets Targeting ISIS*[2] is a dataset intended to work as a counterpoise to the *How ISIS Uses Twitter* dataset. The dataset was gathered over two separate days (7.4.2016 and 7.11.2016) and consists of over 122K tweets from around 96K distinct users, many of which are blocked as of today. The tweets were collected based on the following keywords:

- isis
- isil
- daesh
- islamicstate
- raqqa

[1] `https://www.kaggle.com/fifthtribe/how-isis-uses-twitter`

[2] `https://www.kaggle.com/activegalaxy/isis-related-tweets`

- mosul
- Islamic State

The dataset is not considered to be a perfect counterpoise by the author, as it contains a number of pro-ISIS fanboy tweets. However, it is considered to provide a backdrop against the *How ISIS Uses Twitter* dataset. As opposed to the *How ISIS Uses Twitter*, this dataset has additional information on tags, but lack information about location, followers, number of tweets and description.

5.1.5. Fernandez et al. (2018)

As part of Fernandez et al. (2018) study to understand the roots driving individuals to radicalisation, they leveraged the *How ISIS Uses Twitter* and *Tweets Targeting ISIS* datasets to build a complete dataset of tweets labelled as pro-ISIS and neutral ISIS. Their study was concerned with predicting and detecting the level of radicalisation influence that a user is exposed to. Their data collection approach was grounded in social science theory, yielding a model which they referred to as *The Roots of Radicalisation.* The roots of radicalisation consists of three driving factors for radicalisation, namely:

- **Micro factors:** Factors that self-affect individuals, such as perceived deprivation, injustice, and types of threat that can cause individuals to seek out radical groupings.
- **Meso factors:** Factors related to group thinking, causing individuals to seek out like-minded groups to find support for their ideas.
- **Macro factors:** These roots of radicalisation are related to the influence of government and society, that poses a threat to the group's identity.

Based on the notion of these roots, tweets were labelled as follows:

- Micro roots were captured by tweets created by a user.
- Meso roots were captured by the tweets that a user shared on Twitter.
- Macro roots were captured by the links posted by users as part of their tweets.

The tweets were labelled by considering the posts of each user as a vector of the micro and meso influence that a user is subjected to. These vectors were then broken down into smaller units of n-grams. The frequency of the n-grams was then computed as the frequency of n-gram in the post, normalised by the number of tweets posted by the user. The tweets were then labelled based on the hypothesis that if any of the vectors contained radicalised terminology, then there is considered to be a certain radical influence over the user. The dictionary of radical terminology was based on the authors collecting and extending existing dictionaries of radicalised terminology. The final dictionary was comprised of the following dictionaries:

- ICT Glossary [3]
- Saffron Experts [4]
- Saffron Dabiq Magazines
- Dictionary generated by Rowe and Saif (2016)

The annotation scheme of Fernandez et al. (2018) is similar to that of Rowe and Saif (2016), but instead of using the original words of the tweets, Fernandez et al. (2018) split the words into n-grams. As a counterpoise to the first dataset Fernandez et al. (2018) used the *Tweets Targeting ISIS* dataset. To ensure that the dataset only contained non-pro-ISIS users, the authors randomly selected 40 accounts that were still active at that point. The accounts were manually verified by two annotators to reach a consensus on the labelling of each account.

5.1.6. Ferrara et al. (2016)

Ferrara et al. (2016) carried out a study with the aim to detect extremist users and estimate whether regular users would adopt extremist content. The study was dependent on data and labels constructed and verified by experts in Arabic language. To leverage this dependency, the authors retrieved a dataset of 25K Twitter users labelled as pro-ISIS by the *Lucky Troll Club* crowd-sourcing initiative. All users in the dataset had been suspended on Twitter due to showing support of the Islamic State group. For each account, the dataset contained information about the suspension date and the number of followers. This information was used to collect information not only about the suspended accounts, but also create a dataset of their targets. As the initial dataset consisted of suspended accounts, the authors retrieved missing data by leveraging a dataset previously collected by the Indiana University Davis et al. (2016). This data was collected using the Twitter *gardenhose* data source which contains roughly 10% of the Twitter data stream. After the final processing, the authors were left with two datasets:

- ISIS Accounts: a dataset of around 3,4M tweets from the 25K users in the initial dataset.
- Users exposed to ISIS: a dataset containing around 29,2M tweets generated by around 25K users randomly sampled from the followers of the *ISIS Accounts* dataset.

5.2. Datasets for Personality Prediction

Several studies have aimed to automatically predict the personalities of social media users by using publicly available information found on the web. These studies have spiked an interest in understanding the underlying human traits that can be revealed

[3] `https://www.ict.org.il`

[4] `http://www.saffron-project.eu/`

from social media activity and in turn generated several datasets that can be used for such predictions. This section will go through some of the most common datasets used within the study of automatic personality prediction.

5.2.1. myPersonality Dataset

Among the datasets used in studies on automatic personality prediction, myPersonality is by far the most used dataset and has had tremendous impact on the research field. The dataset was generated through the Facebook application, myPersonality[5]. The application was developed in 2007 by David Stillwell, and in 2009 Michal Kosinski joined the project. The application allowed Facebook users to take a questionnaire and receive scores on their personality traits. The scores were based of the *Big5 Personality Model*. Users were then free to share their results, along with Facebook data, to aid researchers in understanding how personality traits unveil themselves on social media. According to the creators, more than 6 million people participated in the survey, where around 40% of them chose to donate their data. The project closed in 2012, and in 2018 the creators stopped sharing the data with scholars, due to lack of time needed for maintaining the dataset and keeping up with everchanging privacy regulations.

The dataset consisted of Facebook user from various age groups, background and cultures and offer a rich and unbiased view on the personality of social media users. In addition to the personality scores of each user, the dataset contained demographic information, recordings of social media behaviour, preferences, interests, opinions, and much more.

5.2.2. Twitter Myers-Briggs Personality Type Dataset

The Myers-Briggs Personality Type Dataset[6] is a dataset collected through the *Personality Cafe Forum*[7]. The dataset consists of around 8,600 Twitter users along with their Myers-Briggs Personality Type (MBTI) scores. The scores consist of the users four-letter MBTI type, which yield one of 16 different personality types. In addition to the type of the user, the dataset contains the 50 latest posts of each user. The dataset only contains one column, with each row representing a user with its corresponding data. The row of the dataset has the following structure:

- **Type**: The users four-letter MBTI type.
- **Tweet**: The 50 latest tweets by the user, each separated by "|||".

MBTI has been called into question recently due to the unreliability of the experiments surrounding it. Nevertheless, the Twitter MBTI dataset is still being clung to today, as it offers insight into patterns, types and styles of writing provided by social media users.

[5] https://www.psychometrics.cam.ac.uk/productsservices/mypersonality

[6] https://www.kaggle.com/datasnaek/mbti-type

[7] https://www.personalitycafe.com/about/

5.2.3. PAN 2015 Author Profiling Task Dataset

The 2015 Author Profiling Task dataset[8] is a dataset consisting of English, Spanish, Dutch, and Italian tweets, annotated with Big5 personality scores (both binary and continuous values) on a user level. The dataset was published as part of the PAN 2015 Author Profiling task, originally a task for predicting the authors demographics. The dataset has later been used in several personality prediction studies. The personality scores are given in the range of -0.5 to 0.5 and are generated through *Short Big Five Test (BFI-10)*, a Big5 test consisting of ten questions that are meant to capture the underlying personality of a person. Each entry in the dataset is identified by a unique author-id, with corresponding data for the given author. In addition to the tweets and personality trait scores, the dataset contains the following information for each author:

- **gender**: the gender of the author, either male or female.
- **age_group**: the age group of the author, split into groups of 18-24, 25-34, 35-49 and 50-.
- **language**: The language of the author, given as *en* (English), *es* (Spanish), *nl* (Dutch) or *it* (Italian).

Due to privacy, the Twitter handle and UserID are removed for each of the users and replaced by a unique ID for each of the users.

5.2.4. Stream-of-Consciousness Essay Dataset

The *stream-of-consiousness*[9] dataset consist of 2468 essays written by students. The dataset was published by Pennebaker and King (1999) as part of their study on the languages reflection of personality types. Each of the essays are annotated with binary labels (y/n) for each of the Big5 personality traits. The traits are found by each student taking a self-reporting questionnaire. The dataset consists of seven columns, with each row representing each student, and is given on the following format:

- **#AUTHID**: A unique identified for each author/student.
- **TEXT**: The essay written by a student.
- **cEXT**: The Extraversion big5 personality type, with the corresponding binary label *y* (yes) or **n** (no).
- **cNEU**: The Neuroticism big5 personality type, with the corresponding binary label *y* (yes) or **n** (no).
- **cAGR**: The Agreeableness big5 personality type, with the corresponding binary label *y* (yes) or **n** (no).

[8] https://pan.webis.de/clef15/pan15-web/author-profiling.html

[9] https://drive.google.com/file/d/1bbbn8kSBmcVObafdzAQEipRBc4SVqwtb/view

- **cCON**: The Conscientiousness big5 personality type, with the corresponding binary label y (yes) or **n** (no).
- **cOPN**: The Openness big5 personality type, with the corresponding binary label y (yes) or **n** (no).

6. Related Work

This chapter will describe the research methods conducted to identify the state-of-the-art within the field of personality prediction and identifying radicalisation in social media. The chapter starts off by describing the proposed research method and the process of conducting a structured literature review, followed by research on preprocessing and features. Finally, methods used for personality prediction and identifying radicalisation in social media are presented. The majority of the sections included in this chapter were written as part of a preliminary study for on radicalisation and personality prediction.

6.1. Structured Literature Review

A structured literature review was conducted to form the basis for understanding the field of personality prediction and identifying radicalisation in social media, along with identifying the current state-of-the-art. The motivation behind choosing this approach was to gain unbiased and sufficient knowledge in the field. The structured literature review conducted in this project is based on the method proposed by Kofod-Pedersen (2018), which consists of three steps: (1) planning, (2) conducting and (3) reporting. To ensure reproducbility, each step of the structured literature review is presented in the review protocol. Papers gathered outside the scope of the structured literature review were based on recommendations from the project supervisor Björn Gambäck and on citations in papers gathered from the initial query in the structured literature review.

6.1.1. Planning the Structured Literature Review

As part of the structured literature review, research questions were formulated. The research questions were formed on the basis of the initial research goal for the project and are given in Section 1.2.

6.1.2. Conducting the Structured Literature Review

The structured literature review proposed by Kofod-Pedersen (2018) is a five-step process. Each step is described in detail below.

Step 1: Identification of Research
The first step of conducting a structured literature review is specifying the search domain and defining search terms. The search terms are specified in Table 6.1. The search

Table 6.1.: Search Terms

	Group 1	Group 2	Group 3	Group 4
Term 1	Extremism	Machine Learning	Social Media	Personality Prediction
Term 2	Radicalisation	Computational	Twitter	Personality Recognition
Term 3	Terrorism	Detection	Facebook	Personality Profiling
Term 4		Prediction		Personality Profiling

terms presented in Group 1 are concerned with detecting research related to predicting radicalisation, while Group 4 is related to personality prediction. Group 2 is meant to restrict the query to computational approaches, while Group 3 restricts the query to research related to social media. Google Scholar was chosen as the main search domain for the project due to the search engine's ability to retrieve papers from multiple sources while offering ranking and filtering for the queries. As the project is concerned with both detection of radicalisation and prediction of personality, two separate queries were formulated based on the search terms.

1st Query: `(Extremism OR Radicalisation OR Terrorism) AND (Machine Learning OR Computational OR Detection OR Prediction) AND (Social Media OR Twitter OR Facebook)`

The first query string returned a total of 29,700 papers. Due to the rapid change in the state-of-the-art within the field of computer science, the publication period was restricted to 2011-2021. As the main objective of this project is related to personality prediction and identifying characteristics of individuals, only the first 40 highest ranking papers were selected for assessment according to the inclusion criteria.

2nd Query: `(Machine Learning OR Computational OR Detection OR Prediction) AND (Personality prediction OR Personality recognition OR Personality detection OR Personality profiling) AND (Social Media OR Twitter OR Facebook)`

The second query string returned a total of 18,400 papers for the publication period 2011-2021. The first 70 highest ranking papers were selected for assessment according to the inclusion criteria. A higher number of papers were selected for the second query as reviewing literature related to identifying characteristics of individuals is considered to be the main objective of the project.

Step 2: Selection of Primary Studies

From the two query strings, a total of 110 papers were returned. The title and abstract of each paper were assessed according to the primary inclusion criteria found in Appendix A. Examples of these inclusion criteria are *The study is a primary study presenting empirical results* and *The study's main concern is prediction of personality.* The papers passing the first primary inclusion criteria were passed over for assessment according to the secondary primary inclusion criteria. An example of these inclusion criteria is *The study describes an implementation of an algorithm for predicting personality.* After assessment the remaining papers were reduced to a set of 25 papers.

Step 3: Quality Assessment

The 25 papers passing the secondary inclusion criteria were passed on for quality assessment and awarded scores according to the quality assessment criteria presented in Appendix B. Examples of these criteria are *Is there a clear statement of the aim of the study* and *Is the study put into the context of other studies and research.* For each of the criteria, 1 point was given for full fulfillment, 1/2 point for partially fulfilled, and 0 points if the criteria were not met. After the quality assessment another two papers were dropped due to low scores.

Step 4: Data Extraction

Data was extracted from the primary studies according to the list presented in Appendix C and Appendix D.

Step 5: Data Synthesis

The data collected from the data extraction step is presented in Table 3.2 and Table 3.3.

6.1.3. Results of the Structured Literature Review

The results of the initial literature review and structured literature review can be found in Table 3.2 and Table 3.3. Papers reviewed post the initial literature review are not included in the table, but can be found in the bibliography. The rest of the chapter is dedicated to presenting the findings from the reviewed literature.

Table 6.2.: Research on Automatic Personality Prediction Retrieved from the Initial Literature Review

ID	Author(s)	Title	Year	Personality Model	Algorithm	Features	Dataset	Relevant Findings and Conclusion
1	Lima & Castro	A multi-label, semi-supervised classification approach applied to personality prediction in social media	2014	Big 5	NB, SVM, MLP	LIWC, MRC	Twitter (Obama-McCain Debate (OMD), Sanders, & Sem-Eval2013)	Approximately 83% accuracy, with some personality traits classified more accurately than others
2	Tandera, Suhartono, Wongso, & Prasetio	Personality Prediction System from Facebook Users	2017	Big 5	NB, SVM, LR, GB, LDA, Deep Learning	LIWC, SPLICE, SNA, Word Embeddings	Facebook (myPersonality and manually collected data)	Experiments show that deep learning can improve accuracy of personality prediction even when the accuracy is low for some traits.
3	Liu, Perez, & Nowson	A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts	2016	Big 5	RNN	Character-to-word-to-sentence	Twitter (PAN 2015 Author Profiling Task Dataset)	State-of-the-art results on the user lever, and the model performs reasonably well on short texts.

4	Park, Schwartz, Eichstaedt, Kern, Kosinski, & Stillwell	Automatic Personality Assessment Through Social Media Language	2015	Big 5	Regularized Ridge Regression	Words, phrases, and Topics	Facebook (mypersonality)	Social Media Language can be used to create valid and reliable measures of personality
5	Farnadi, Zoghbi, Moens, & De Cock	Recognising Personality Traits Using Facebook Status Updates	2013	Big 5	NB, SVM, kNN	LIWC	Facebook (myPersonality) and annotated Essays	Beat baseline, even with small data samples. Personality prediction generalises well across domains.
6	Arnoux, Xu, Boyette, Mahmud, Akkiraju, & Sinha	25 Tweets to Know You: A New Model to Predict Personality with Social Media	2017	Big 5	Gaussian Processes Regression	Word Embeddings	Twitter (manually collected data)	Outperform state-of-the art with 8 times fewer input data.
7	Golbeck, Robles, & Turner	Predicting Personality with Social Media	2011	Big 5	M5'Rules, Gaussian Processes	LIWC	Facebook (manually collected)	Predicted each personality trait to within 11% of its actual value.
8	Bachrach, Kosinski, Graepel, Kohli, & Stillwell	Personality and Patterns of Facebook Usage	2012	Big 5	Linear Regression	Number of likes, photos, status updates, tags, & friends	Facebook (myPersonality)	Personality traits correlated with patterns of Facebook usage

9	Preotiuc-Pietro, Carpenter, Giorgi, & Ungar	Studying the Dark Triad of Personality through Twitter Behaviour	2016	The Dark Triad	Linear Regression	LIWC, Unigrams, Word Clusters, Profile Pictures, Sentiment, Emoticons, Shallow Textual Features	Twitter (manually collected)	Reliable accuracy with Pearson Correlation of around 0.25.
10	Kumar & Gavrilova	Personality Traits Classification on Twitter	2019	MBTI	Ensemble Gradient Boosting Decision Trees, SVM	TF-IDF & Word Embeddings (GloVe)	Twitter (Twitter MBTI Personalit Dataset)	Simple language-models can reliable estimate certain personality traits, but do not sufficiently capture all personality traits due to the limited format on Twitter
11	Golbeck, Robles, Edmondson, & Turner	Predicting Personality from Twitter	2011	Big 5	Gaussian Prosesses, ZeroR	LIWC	Twitter (manually collected)	Predicted personality trait to within 11-18% of the actual value.

12	Peng, Liou, Chang, & Lee	Predicting Personality Traits of Chinese Users Based on Facebook Wall Posts	2015	Big 5	SVM	Bag-of-Words, TF-IDF	Facebook (manually collected)	Accuracy of 73.5% for extraversion. Using a language specific tokeniser improves precision by up to 60%.
13	Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, Agrawal, Shah, Kosinski, Stillwell, Seligman, & Ungar	Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach	2013	Big 5	Ridge Regression	Words, phrases and topics	Facebook (myPersonality)	Using an open-vocabulary for feature extraction supersedes the results obtained using LIWC features.
14	Quercia, Kosinski, Stillwell & Crowcroft	Our Twitter Profiles, Our Selves: Predicting Personality with Twitter	2011	Big 5	M5’ Rules Regression	Followers, Following, & Lists	Twitter (manually collected)	Personality can be easily and effectively predicted from public data.

15	Sumner, Byers, Boochever, & Park	Predicting Dark Triad Personality Traits from Twitter usage and a linguistic analysis of Tweets	2012	The Dark Triad	Ensemble (SVM, RF, NB, C4.5 Decision tree)	LIWC	Twitter (Collected trough questionnaire)	Clear relationship between Twitter activity and Dark Triad personality traits. Poor performance when the model is applied directly to individuals. Need for standardised evaluation metrics in the field
16	Carducci, Rizzo, Monti, Palumbo, & Morisio	TwitPersonality: Computing Personality Traits from Tweets Using Word Embeddings and Supervised Learning	2018	The Big 5	SVM, Linear Regression, LASSO	Word Embeddings	Twitter (Using myPersonality as benchmark)	Treats Twitter posts individually. Achieves good conversion for the models, but they seem to lack discriminative power.
17	Plank & Hovy	Personality Traits on Twitter	2015	MBTI	Linear Regression	Binary word n-grams and discretised count-based meta features	Twitter (manually collected)	Model I-E and F-T fairly well, struggles with other personality types.

18	Xue, Wu, Hong, Guo, Hao, Wu, Zhong, & Sun	Deep Learning-Based Personality Recognition from Text Posts of Online Social Networks	2018	Big 5	AttRCNN, CNN, GBR, RF, MLP, SVR	LIWC, Word Embeddings	Facebook (myPersonality) & Twitter	Combining deep semantic features with statistical linguistic features and feeding the feature vectors to traditional regression algorithms yield the lowest average MAE.

Table 6.3.: Research on Identification of People at Risk of Radicalisation Retrieved from the Initial Literature Review

ID	Author(s)	Title	Year	Algorithm	Features	Dataset	Relevant Findings and Conclusion
19	Roew & Saif	Mining Pro-ISI Radicalisation Signals from Social Media Users	2016	-	BoW	Twitter (manually collected)	Social dynamics play a strong role in adopting extremist terms. The adoptions of extremist terms increase significantly prior to being radicalised.
20	Araque & Iglesias	An Ensemble Method for Radicalisazation and Hate Speech Detection Online Empowered by Sentic Computing	2021	Logistic Regression, SVM	TF-IDF, Word Embeddings (SIMON), AffectiveSpace, SenticNet	Twitter (Pro-Neu, Pro-Anti, & Magazines)	Adding affect features to domain textual representations in an ensemble increases classification performance.
21	Asif, Ishtiaq, Ahmad, Aljuaid, & Shah	Sentiment Analysis of Extremism in Social Media from Textual Information	2020	NB, SVM	TD-IDF, Bag-of-n-grams	Facebook (manually collected)	Able to predict different levels of extremism with an average F1-score of 0.81

22	Fernandez, Asif, & Alani	Understanding the Roots of Radicalisation on Twitter	2018	Collaborative Filtering	n-grams (terms & expressions), Lexical Features	Twitter (Kaggle & manually collected)	F-score of 0.906 for detecting radicalisation and MAE of 0.1025 for predicting risk of radicalisation.
23	Ferrara, Wang, Varol, Flammini, & Galstyan	Predicting Online Extremism, Content Adopters, and Interaction Reciprocity	2016	RF, Logistic Regression	User Metadata, Timing Features, Network Statistics	Twitter (Lucky Troll Club pro-ISIS list)	Obtained fair results for six forecasting combinations, with AUC ranging from 72% to 93%.
24	Agarwal & Sureka	Using kNN and SVM-Based One-Class Classifier for Detecting Online Radicalization on Twitter	2015	SVM, kNN	Term Frequency	Twitter (UDI-TwitterCrawl-Aug2012 & ARM-TwitterCrawl-Aug2013)	Obtain best results for SVM classifier with F-score of 0.83.
25	Benigni, Joseph, & Carley	Online Extremism and the Communities that Sustain it: Detecting the ISIS Supporting Community on Twitter	2017	SVM & RF	User metadata, hastags	Twitter (manually collected)	Approach outperforms several of the arroches on classification tasks for identifying ISIS supporters.

6.2. Datasets and Extraction of Data

Chapter 5 gave a presentation of some existing popular datasets used for personality and radicalisation studies. While some studies use existing datasets for their studies, others expand upon these datasets or create new ones. This sections gives an overview of the usage of existing data within related work.

6.2.1. Extremism and Radicalisation

Identifying radicalisation has been a subject of increased attention within social science research. Social media plays a central role in online extremism, propaganda and efforts of radicalisation (Fisher, 2015). Social media platforms offer vast amounts of data for analysing extremist environments and identifying the trajectories to radicalisation. Still there remains a great need for a gold standard dataset to train models for detection (Ferrara et al., 2016). Still, extremist and radicalisation research vary in their goals. For this reason, manually collecting data or expanding upon existing dataset seems to be the preferred approach.

Section 5.1 gave an overview of some pre-existing datasets within radicalisation research. Of the reviewed work, the *How ISIS Uses Twitter* seems to be the most used dataset. In their study of measuring radicalisation risk on social media, Lara-Cabrera et al. (2017) used the *How ISIS Uses Twitter* to identify common behaviour patters between users contained in the dataset by measuring similarities among the users with regards to a set of defined indicators of ongoing radicalisation. Among their findings were that the users shared a feeling of being discriminated for being Muslims and expressed a negativity towards Western society. Araque and Iglesias (2022) used a combination of three datasets from previous research for detecting radicalisation online, namely the Pro-Neu dataset (Fernandez et al., 2018), Pro-Anti (Rowe and Saif, 2016), and the Magazines dataset (Araque and Iglesias, 2020). Two of these datasets were presented as part of Section 5.1. Their study used of ensemble learning methods for classifying users as either radical or not. The datasets were used separately with the models. The highest score was achieved for the dataset of Fernandez et al. (2018), with a F1-score of 98.21. The dataset of Rowe and Saif (2016) achieved a best score of 90.90, while Araque and Iglesias (2020) scored 94.66.

While several radicalisation studies built their datasets upon existing datasets, Agarwal and Sureka (2015) built their dataset of radical users from two datasets of randomly sampled Twitter users, namely the UDI-TwitterCrawlAud2012[1] and ATM-TwitterCrawl2013[2]. Their study focused on classifying users as either radical or not, using KNN- and SVM-based one-class classifiers. Radical users were identified through a semi-supervised approach, making use of a small set of users labelled as radical

[1] `https://wiki.illinois.edu//wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012`

[2] `https://wiki.illinois.edu/wiki/display/forward/Dataset-ATM-TwitterCrawl-Aug2013`

and a bigger set of unlabelled users. The labelled users were identified using a set of predefined hashtags, before being manually verified by the annotators.

6.2.2. Personality Prediction

As described in section 2.1, Social Media offer a vast amount of data, that when utilised, can offer insight into the habits of its users. Most larger Social Media platforms today offer access to this data, either through established datasets, such as Facebook's myPersonality Project or Twitter's Application Programming Interface (API). Research in the field of personality prediction have used both existing datasets and manually collected and annotated data. The approaches differ between researchers, but the resulting datasets often share common characteristics.

When it comes to personality prediction from Facebook data, the Facebook application *myPersonality* has had tremendous impact on research in the field. The myPersonality Facebook dataset was described as part of Section 5.2. In an effort to create a Deep Neural Network for processing feature vectors to be used for personality prediction, Xue et al. (2018) used a subset of 115,864 users from the myPersonality Project, containing a total of 11,494,862 status updates. Carducci et al. (2018) used the same dataset with a subset of 250 users and 9,913 status updates as a benchmark for an in-house transfer learning model for personality prediction on tweets. In one of the largest studies, by an order of magnitude, of language and personality, Schwartz et al. (2013) collected a dataset of over 15.4 million Facebook messages from 75,000 volunteers as part of the myPersonality Project. The data was used as part of a descriptive study of the correlations between written language and gender, age, and personality. Bachrach et al. (2012) used the dataset to train a model capable of detecting correlations between personality traits and patterns of Facebook usage and Park et al. (2015) used the same data in an effort to create a valid and reliable model for measuring personality. Tandera et al. (2017) and Farnadi et al. (2021) used the dataset in combination with other data from other domains to train predictive models for personality and found that data from the myPersonality project generalised well across domains. Tandera et al. (2017) used a combination of manually collected data and established datasets. In their study they combined two datasets where the first one contained 250 users with around 10,000 status updates obtained from myPersonality, while the second consisted of 150 users which were collected manually. The manually collected set was fed into Apply Magic Sauce [3] application which is developed by Cambridge Psychometrics Centre to predict psychological traits from digital footprints of human behaviour. Quercia et al. (2011) used the myPersonality dataset to gather information on the Twitter accounts of Facebook users contained in myPersonality. Using this approach, they were able to build a dataset consisting of 335 Twitter users. The dataset was expanded further with personality data, such as number of followers and whether the users had been listed in

[3] `https://applymagicsauce.com/about-us`

others' reading lists. Building upon the original myPersonality dataset, they were able to achieve a RMSE-score of 0.88 across the big5 traits.

In addition to the myPersonality dataset, Section 5.2 presented several Twitter datasets which are used within personality research. Kumar and Gavrilova (2019) used the Twitter MBTI Personality Dataset for classifying MBTI personality traits of Twitter users using Ensemble Gradient Boosting Decision Trees and Support Vector Machines. The best result was obtained for the ensemble model, with an average F1-score of 76.5 across the traits, yielding the highest score for Sensation-Intuition of 0.92. In their study on recognising personality traits from short texts, Liu et al. (2017) used the English, Spanish, and Italian parts of the PAN 2015 Author Profiling Task Dataset [4]. They comprised their own implementation of a bidirectional RNN, yielding state-of-the-art results across all five traits and three languages.

Despite the popularity of pre-existing dataset, many researchers choose to collect their own data. Manually collecting and annotating data enables the researchers to customise queries and retrieve more domain specific data. In their study on predicting personality traits from Facebook wall posts, Peng et al. (2015) designed a short Big 5 questionnaire based of the Big Five Mini-Makers and collected a dataset of 222 Facebook users who had Chinese as their main written language. The dataset consisted of user metadata and status updates, along with the personality traits of each user. They found that comprising their own dataset enabled the inclusion of side information which in turn improved upon their results. In their study of personality prediction on Facebook, Golbeck et al. (2011b) developed their own Facebook application to manually collect data from 279 individual Facebook users. The application consisted of a 45-question version of the Big 5 questionnaire which allocated personality trait scores to each participating user, while also collecting profile information. By collecting their own dataset, they were able to comprise a total of 74 features per user, which resulted in prediction within 11% of the actual trait score. In a similar study on personality prediction from publicly available Twitter data, Golbeck et al. (2011a) used their 45-question version of the Big 5 Personality Inventory to collect the personality traits of each participating user, their most recent 2,000 tweets, their number of followers, and number of followings. For this study, they were able to predict the personality traits of the twitter users within 11%-18% of the actual value. Carducci et al. (2018) has a selection of Twitter users take the Big5 Inventory Test and used the Twitter API to scrape information on these users. The test set was compared to a gold standard of Facebook status updates gathered from the myPersonality project. The results of their study showed lower MSE scores for their own comprised dataset than that of the myPersonality dataset. Similar to that of (Carducci et al., 2018), Arnoux et al. (2017) built their Twitter dataset of users who agreed to share their Twitter data and take 50-item IPIP to measure their Big 5 personality traits. Using a combination of word embeddings and regression they were able achieve comparable or better accuracy than state-of-the-art techniques with a

[4] `https://pan.webis.de/clef15/pan15-web/author-profiling.html`

sample of only 1,300 users. Sumner et al. (2012) used a similar approach when trying to predict the Dark Triad personality traits of users on Twitter. They recruited nearly 3,000 users to answer a questionnaire for predicting Dark Triad personality traits. The questionnaire was administered as a Twitter application that collected up to 3,200 tweets from each user, along with user meta data. They found their prediction models to be unsuitable on an individual level, but still able to provide important insight to anti-social behaviour for larger samples of users. Similar to that of Sumner et al. (2012), Preotiuc-Pietro et al. (2016) built their own dataset by having Twitter users take the Dirty Dozens test, a questionnaire to assess the Dark Triad personality traits of an individual. The dataset was filtered for users with more than 500 tweets, resulting in 536,579 tweets from 491 distinct users. As opposed to Sumner et al. (2012), they found their prediction model to produce reliable accuracy for out-of-sample users.

Plank and Hovy (2015) collected 1.2 million tweets from 1,500 distinct users on Twitter. Their approach differed from the ones mentioned above, as their dataset was collected by searching for users that *self-identified* with one of the 16 MBTI's. The users were identified by searching for mentions of any of the 16 MBTI types, in addition to the term "Briggs". Tweets containing more than one MBTI type or were the gender of the user could not be verified were discarded. Using this collection approach, the results showed that they were able to model the Introvert-Extrovert and Feeling-Thinking distinction fairly well, while other types proved more difficult to model.

6.3. Preprocessing and Feature Selection for Detecting Radicalisation

This section presents the steps taken in preprocessing data and the related feature selection to be used to detecting signs of radicalisation and predicting social media users prone to extremism.

6.3.1. Preprocessing

According the literature review, preprocessing data used for detecting signs of radicalisation and predicting users prone to extremism rely to a large degree on social science research. Rowe and Saif (2016) mined for radicalisation signals from Twitter users. Their approach relied completely on identifying lexical terms used within the classes of pro-ISIS and anti-ISIS users. Users were classified according to the micro, meso and macro factor model described in Section 2.2.1. Due to their lexical approach, all text was considered important. Little preprocessing was applied, except for removing stop words. Araque and Iglesias (2022) used datasets from previous research on radicalisation detection (Fernandez et al., 2018; Rowe and Saif, 2016). In order to comply properly with the semantic tools utilised, words were lemmatised for Affective Space and SenticNet, and one-hot encoding applied to represent categorical features for the SenticNet tool. Additionally, images, links and other were removed, leaving only the text. Asif et al. (2020) base their preprocessing on the workings of Winkler (2003), a guide for preprocessing data. Among the preprocessing steps proposed, they chose to remove URLs, emojies, convert words to lowercase, and combine all data in one spreadsheet. Fernandez et al. (2018) used a similar approach to that proposed in the guide, by removing URLs, numeric and punctuation symbols, stop words and infrequent n-grams.

Not all preprocessing is related to removing terms or cleaning up instances, but rather removing the whole instance. Benigni et al. (2017) main preprocessing task was related to identifying and removing users from their dataset who were not pro-ISIS, but rather following pro-ISIS for other reasons, such as news outlets. Their preprocessing was mainly concerned with removing irrelevant instances from the dataset, rather than cleaning the included instances.

6.3.2. Feature Selection

Features used for detecting signs of radicalisation and predicting social media users prone to extremism consist of both linguistic features from the users' written text and user specific features such as metadata and profile information. This section presents the feature selection identified by the literature review.

Textual Features

The methods for extracting textual features vary among the reviewed research, depending much upon the research goal at hand. Agarwal and Sureka (2015) used Term Frequency as their main feature for their kNN model and Support Vector Machine. A common approach, found in the reviewed literature, for extracting and evaluating important words is the Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is often used in combination with other textual features, as TF-IDF makes no use of semantic similarities between words. Araque and Iglesias (2022) used TF-IDF in combination with Word Embeddings, AffectiveSpace and SenticNet for extracting semantic features from tweets. Their approach was not concerned with obtaining complete representations of the text, but rather computing affect aware features. Asif et al. (2020) used TF-IDF in combination with Bag-of-n-Grams, choosing not to consider the semantic similarities of words. Their study was mainly concerned with analysing sentiments of extremists in social media, a task where TF-IDF proved useful by assigning weight to important words.

User Specific Features

User specific information such as metadata, profile information and number of status updates, are often used as complementary features to textual features, but can also be utilised alone. Benigni et al. (2017) used hashtags and profile specific information from Twitter users such as creation date, tweet count, and count of followers as their features. The study was concerned with analysing the Twitter communities of online extremists and used user specific features for analysing the networks of users categorised as extremists.

Ferrara et al.'s (2016) research aimed at detecting potential content adopters of online extremism and the point of activation for extremist behaviour. For this they extracted a number of features related to the Twitter users; (1) user metadata features (followers, posted tweets, mentions), (2) temporal features (average number of tweets per day and the time interval between two consecutive tweets), and (3) Network Statistics (distribution of retweeters' number of followers and distribution of mentioners' number of friends). These features were used to create a network for the information flow between users and determine the point of extremist content adoption.

6.4. Preprocessing and Feature Selection for Personality Prediction

This section presents the steps taken in preprocessing data and the related feature selection to be used to personality prediction.

6.4.1. Preprocessing

The common steps of preprocessing data for personality prediction are presented in Section 4.1.1 and include converting words to lowercase, stemming, tokenisation, and stop words removal. The degree to which these preprocessing steps are utilised vary among studies and often depends on the end goal of the study. Social Media language is considered to be more informal than other written text and thus noisier. Therefore, preprocessing often involves the task of balancing between removing noise and keeping information (Liu et al., 2017). To maintain this balance, preprocessing often requires some domain specific knowledge language about the writers. Xue et al. (2018) found that noisy and informal language can contain meaningful semantic information, but might also worsen the quality of word embeddings, and in turn, the quality of the personality prediction model. They found that there is a difference in meaning when a word is written in its original form versus when elongations are used. As an example, the study found that the use of the words *"good"* and *"gooooood"* inherit different semantic meanings. To cope with this semantic difference, the length of these tandems was reduced to tokens of size three, keeping the original semantic meaning expressed through the tandem. Carducci et al. (2018) found that the use of stemming on social media language resulted in semantic meaning being lost.

Language in social media often involves the use of unconventional characters like # for marking an update with a subject, @ for mentioning a specific user or URLs for referring to another web-page. Liu et al. (2017) found that normalising text by mapping hashtags, URLs and mentions to standardised abbreviations yielded better results for their classifier by reducing the risk of modelling character usage not directly linked with personality. The count of hashtags, mentions and URLs was used as features for the model. Arnoux et al. (2017) took this a step further and chose to remove all hashtags and URLs completely from their data. Carducci et al. (2018) used a similar approach by removing URLs and mentions from their data. Removing these unconventional characters should however be considered carefully. Golbeck et al. (2011b) found that users actively sharing URLs had a higher degree of openness than users who did not. This suggests that important information may be kept by mapping URLs to standardised abbreviations. Similarly, Sumner et al. (2012) found, when analysing Dark Triad personality traits on Twitter, that the use of hashtags and mentions were significantly correlated with narcissistic traits.

Liu et al. (2017) chose a completely different approach in their deep learning personality prediction study. As their study was working with a dataset of multiple languages, they proposed a feature engineering-free modelisation, making their model language independent. However, they did apply tokenisation to the tweets, and normalised each text by mapping URLs and mentions to single characters.

6.4.2. Feature Selection

Features used for personality prediction consist of both linguistic features from the users' written text and user specific features such as metadata and profile information. This section presents the feature selection identified by the literature review.

Textual Features

Facebook and Twitter allow users to share textual content with others, either through posting content or writing short biographical about themselves. The processes of generating features from this textual information may vary. Linguistic Inquiry and Word Count (LIWC) is a gold standard for analysing textual information and is often applied to generate features from written data. As described in Section 4.1.2, LIWC takes written data as its input and counts the percentage of different words that reflect certain emotions, part of speech, thinking styles etc. The written data is then categorised in defined classes such as "sadness", "affect", and "pronouns". As presented in Table 6.1.3 LIWC is among the tools most often applied to extract textual features for personality prediction but is often applied in combination with other textual features or user specific features. Farnadi et al. (2021) used LIWC in combination with user specific features for recognising personality traits from Facebook status updates, while Preotiuc-Pietro et al. (2016) used LIWC in combination with a bag-of-words representation as their features. Xue et al. (2018) used LIWC in combination with Word Embeddings to create a 119-dimensional vector representation of the features. By considering this approach, they were able to capture semantic similarities between words and categories categorised using LIWC.

The popularity of LIWC may be the result of several studies indicating the features aptitude for predicting personality. Golbeck et al. (2011a) found that the LIWC features showed higher average correlation with Big 5 personality traits than the non-LIWC features. Similarly, Tandera et al. (2017) experimented with a combination of open- and closed vocabulary, with LIWC, on two different datasets. They achieved the highest accuracy across both datasets using LIWC without any feature selection. However, applying LIWC on texts for personality prediction may depend upon the domain of the data. Arnoux et al. (2017) compared LIWC features against 3-gram textual feature representation and found LIWC to work better on shorter texts. Social media language differs from other media such as online news outlets or blogs. Social media platforms can impose limits on the number of characters in a status update, resulting in the use of abbreviations, or be more informal, resulting in slang or misspellings. For this reason, a closed-vocabulary approach might not be able to capture relevant categories for predicting personality. Sumner et al. (2012) used LIWC to analyse tweets but found the maximum length of tweets to be a limiting factor. Their study showed that open-vocabulary approaches can yield additional insight and information when predicting personality of Facebook users. Schwartz et al. (2013) used an open-vocabulary approach with words, phrases and topics as features. The motivation behind their approach was that a closed-vocabulary can limit the findings to preconceived relationships with words or categories. The authors claimed that their open-vocabulary analysis yielded further insight into the behavioural residue of personality types. Including LIWC features on top of the open-vocabulary features did not result in any improved accuracy, suggesting that the open-vocabulary approach is able to capture predictive information as well, or even better, than LIWC. Kumar and Gavrilova (2019) share the view on LIWC, and state that it is not clear if such tools can be used directly with social media language, given the dynamic nature and informal language used on social media platforms. Xue et al. (2018) used a deep learning approach to extract deep semantic features, before combining these features with statistical linguistic features and dictionary-based linguistic features and feeding them into a set of different regression algorithms. Their feature analysis showed a lower prediction error for the deep semantic features in comparison to the statistical and dictionary-based features, confirming the statement of Kumar and Gavrilova (2019). Similarly, Mehta et al. (2020) used a transfer learning model in combination with MLP and Support Vector Machines to predict personality traits of people according to the Big 5 model. They extracted two different types of deep features; (1) psycholinguistic features, using Mairesse, SenticNet, NRC Emotion Lexicon, VAD Lexicon, and Readability (based on surface characteristics of the text, namely words, syllables and sentences), and (2) Language Model Features extracted using different variations of BERT. They outperformed the current state-of-the-art results for both their datasets.

User Specific Features

User specific information such as metadata, profile information, and number of status updates, are often used as complementary features to textual features. Facebook allow the users to share a greater number of profile related information than Twitter, resulting in a larger variety of profile related features. User specific features extracted from Twitter rely to a greater extent on the user's individual network. For this reason, the average number of followers, following, tweets, and network density are often used as supplementary features to textual representations (Preotiuc-Pietro et al., 2016; Sumner et al., 2012; Golbeck et al., 2011a). User specific features are proven to have strong correlation with personality traits and many studies therefore include these features in their feature set for personality prediction. Bachrach et al. (2012) extracted a set of user specific features for their personality prediction model, among them, *number of friends*, *number of likes*, and *number of photos*. Their study found strong relationships between features based on Facebook profile information and personality traits. Similarly, Quercia et al. (2011) were able to predict personality traits of Twitter users based purely on the user specific features number of listed counts, followers, and following.

6.5. Identifying Radicalisation

Identifying sign of radicalisation is a subject of substantial interest, ranging from social science research, computational detection research, and national security programs. When tackling the problem of detecting signs of radicalisation and predicting social media users prone to extremism the choice of model varies among the research identified in the literature review.

Most of the algorithms applied in the research use traditional machine learning algorithms. For classification tasks, Support Vector Machines is the most utilised algorithm and is used by Araque and Iglesias (2022); Asif et al. (2020); Agarwal and Sureka (2015); Benigni et al. (2017). Benigni et al. (2017) tested the performance of their Support Vector Machine against the Random Forest algorithm in an effort to detect the ISIS supporting community on Twitter. They found Random Forest to perform better on this task than their Support Vector Machine. Agarwal and Sureka (2015) applied unsupervised learning to the task of detecting radicalisation on Twitter using the k-Nearest Neighbor algorithm. They achieved an F-score of 0.60 for the algorithm, substantially lower than their Support Vector Machine which yielded an F-score of 0.83.

Asif et al. (2020) aimed at predicting the level of sentiment towards extremism among Twitter users. In addition to their Support Vector Machine they applied Naïve Bayes to the problem. They found the Support Vector Machine to outperform Naïve Bayer on the task with an overall accuracy of 82%. Araque and Iglesias (2022) applied Logistic Regression in addition to their Support Vector Machine for detecting sentiments of Twitter users prone to radicalisation. They found their Support Vector Machine to perform best with a Friedman rank of 6.8, in comparison to their best performing Logistic Regression model with a rank of 8.7.

Fernandez et al. (2018) aimed at identifying different roots of radicalisation on Twitter. For this task they applied Collaborative Filtering along with Naïve Bayes, J48, and Logistic Regression. Their best score was achieved for Naïve Bayes, yielding an F-score of 0.9 for detection, and between 0.7 and 0.8 for prediction. Their features were based on the *Roots of Radicalisation* model described in Section 2.2.1, representing one feature for each of the roots. Ferrara et al. (2016) aimed their study at predicting online extremism and identifying content adopters. For this task they used Random Forest and Logistic Regression for modelling the problem. Random Forest proved to be the best algorithm, yielding an F-score of 0.874 and AUC of 0.871. In comparison, Logistic Regression only yielded an F-score of 0.599 and AUC of 0.756.

Rowe and Saif (2016) data mining-oriented approach to radicalisation detection differs from the aforementioned approaches as it utilised no machine learning algorithm. Their approach relied completely on a lexical analysis of tweets using Bag-of-Words. For identifying activation points of users turning to extremism they based their model on two hypotheses grounded in social science research; (1) Sharing incitement material from known pro-ISIS accounts or accounts suspended for supporting ISIS, and (2) Using extremist language synonymous with anti-Western or pro-ISIS rhetoric. Users inheriting one or both of the hypothesis criteria were classified as extremists. Their model was based of the ideas presented in Section 2.2 by depicting radicalisation as a step-wise process, leading to activation of extremist views. The Twitter history of users meeting these criteria were studied, both before and after activation (inheriting hypothesis criteria) to analyse changes in language and sentiment.

6.6. Algorithms for Personality Prediction

When tackling the problem of personality prediction, the choice of algorithm depends much upon the modelling of the problem itself. As mentioned in Section 3.1, machine learning can be treated as either a classification problem where the output is one or several classes, or a regression problem where the output is a continuous value. The various approaches reviewed in the literature are presented below.

Predicting personality can be done with one single algorithm or a combination of different algorithms. In the case of personality classification, the algorithm of choice produces a binary output for a given personality trait. Several algorithms have yielded promising results for personality trait classification. Traditional machine learning algorithms rely on features being manually engineered, as described in Section 6.4.2. Many of these approaches are still used today, although many of the more recent studies are based on deep learning and transfer learning. Tandera et al. (2017); Farnadi et al. (2021); Sumner et al. (2012) used Support Vector Machines and the Naïve Bayes' method separately as their classification model. Tandera et al. (2017) were able to outperform the results of previous studies using the myPersonality dataset for personality classification. Ensemble algorithms have also been used for personality trait classification, where Kumar and Gavrilova (2019) used a combination of Gradient Boosting trees and Support Vector Machines for their model. Different implementations of Decision Trees have also been applied to the classification task, among them Gradient Boosting Trees (Tandera et al., 2017) and Random Forest (Sumner et al., 2012).

When treating the output of personality prediction as a continuous value, regression algorithms are the method of choice. Big 5 is the most applied method for predicting personality, where each personality trait lies on a continuous spectrum. Treating the problem of personality prediction as a regression problem might therefore yield more realistic results. In the early workings of personality prediction Golbeck et al. (2011b); Quercia et al. (2011) used the M5' Rules Algorithm for regression personality prediction. Golbeck et al. (2011a,b); Quercia et al. (2011); Arnoux et al. (2017) all applied the Gaussian Processes algorithm to their regression problem of personality prediction to produce a probability distribution. Several variations of linear regression have also been applied to the problem of personality prediction. Carducci et al. (2018) used LASSO as their regression algorithm for predicting personality, while Park et al. (2015); Schwartz et al. (2013) used ridge regression for their predictive model. Carducci et al. (2018); Preotiuc-Pietro et al. (2016); Bachrach et al. (2012); Plank and Hovy (2015) used the Linear Regression as their predictive model.

With the rise in computational power and advancement in machine learning algorithms, many of today's approaches to personality prediction rely on deep learning methods. Tandera et al. (2017) compared the more traditional machine learning algorithms against a selection of deep learning implementations. They found that deep learning can improve the overall accuracy of personality prediction, even with small datasets and low accuracy for some personality traits. Liu et al. (2017) used a Recurrent Neural Network (RNN) as a predictive model of personality. The motivation behind their approach was the flexibility of the model and its ability to handle previously unseen words. Their model provided state-of-the-art results and performed well with shorter texts, which is an advantage when analysing tweets. Xue et al. (2018) designed a deep learning model called attRCNN consisting of the attention mechanism, RNN and Convolutional Neural Network (CNN). The model was used for extracting deep semantic features before feeding the vectorised features to a Gradient Boosting algorithm. Majumder et al. (2017) applied a similar approach for predicting personality, outperforming the current state-of-the-art for all personality traits of the Big 5. They used a CNN for extracting features before feeding the features into a fully connected neural network. In a more recent study, Mehta et al. (2020) used variations of the BERT algorithm in combination with several implementations of deep learning algorithms to predict personality traits of annotated essays and the Kaggle MBTI dataset. Their best performing model, a combination of BERT-large and MLP, was able to beat the current state-of-the-art on both datasets.

6.7. Statistical Analysis

A large amount of the work carried out prior to applying machine learning relies on analysing statistical relationships between the data and the extracted features. As found by the literature review, a lot of interesting correlations appear between the user profiles, their submitted texts and the extracted features.

6.7.1. Radicalisation

Rowe and Saif (2016) used a purely statistical approach when analysing the radicalisation trajectories of Twitter users and the change in language before and after activation of extremist views. They used Bag-of-Words to extract features from tweets and found clear correlations between the language of the users and the extracted features. Prior to activation, users tended to use words like Syria, Israel and Egypt in a negative context with high frequency. After activation, there appeared to be a clear shift in the language and the activity of the users, where terms like Allah, Muslims, and Quran became more frequent. They also found that the term ISIS was used in a negative context, even though the users were categorised as pro-ISIS. Agarwal and Sureka (2015) found similar results when analysing the feature importance of their Support Vector Machine. They found that the presence of religious terms, war related terms, bad words and negative emotions played an important role among the users categorised as pro-ISIS or radical. Fernandez et al. (2018) analysed the predictive power of their pro-ISIS micro and pro-ISIS meso vector as roots of radicalisation. The micro vector was based on users' perception of deprivation and perceived procedural injustice. The meso vector was related to users' feeling of support from groups and communities that used comparison with other groups to show injustice to create an us-versus-them thinking. These vectors proved to have a clear predictive power for pro-ISIS behaviour.

6.7.2. Personality Prediction

As part Preotiuc-Pietro et al.'s (2016) research on the Dark Triad personality traits of Twitter users, they analysed the inter-correlation of demographic features and word choice with Dark Triad personality traits. They found all Dark Triad personality traits to be significantly correlated with gender and age. Their analysis also found strong correlations between the choice of words from Twitter users and their associated personality traits. Narcissists tended to use words like *favorite* and *things* more often, while psychopaths used words like *injuries* and *women* more frequently. Sumner et al. (2012) also analysed the correlation between The Dark Triad personality traits of Twitter users. Their analysis indicated that users scoring high on machiavellism and psychopathy tended to use more swear words, anger and negative emotions. Users scoring high on narcissistic traits tended to use more terms related to sex, hashtags, and mentions.

Schwartz et al. (2013) analysed the correlations between categories obtained using LIWC and the Big 5 personality traits. Their results showed the LIWC category *Anger* to be predictive of Facebook users scoring low on agreeableness and consciousness. Farnadi et al. (2021) used social network features in addition to LIWC in their study of personality traits of Facebook users. They found the network feature *network size* to be significantly correlated with extraversion, substantiating the intuition of extraverts having a greater social network. Their analysis also indicated *transitivity* having a negative correlation with agreeableness and extraversion. Golbeck et al. (2011b) found similar results in their analysis of the correlation between the Big 5 personality traits of Facebook users and their extracted features, with extraverts having a larger social network. Their analysis also found the use of swear words to be negatively correlated with consciousness, and agreeableness to be correlated with words indicating positive emotion.

7. Data

In order to build a model for predicting personality traits of individuals at risk of radicalisation, it is essential to have both relevant and sufficient data. As mentioned in Chapter 5, the research goal of this Thesis requires the use of multiple datasets. A total of four datasets were acquired for the conducted research. First, a dataset of individuals labelled as being vulnerable to radicalisation was created from a set of radical seed accounts identified manually on Twitter. Second, for the task of personality prediction, a selection of three datasets from Section 5.2 was used for training such a model. This chapter consists of six sections. The first section presents the datasets chosen for training models for personality prediction. The second section gives a thorough explanation of the data collection process conducted for creating a dataset of users at risk of being radicalised. The third section gives an explanation of a manually collected dataset consisting of *not-vulnerable to radicalisation* Twitter users, meant to serve as a counterpoise to the dataset of users vulnerable to radicalisation. The *personality* datasets, *vulnerable to radicalisation* dataset and *not-vulnerable to radicalisation* dataset used in this Thesis all needed different pre-processing steps in order to produce a uniform format. The explanation of the pre-prosessing steps applied were therefore divided three different sections, which make up the last sections of this chapter.

7.1. Personality Datasets

A collection of three datasets were selected to train a model for personality prediction. Two of the datasets consisted of tweets annotated with Big-5 personality traits, while the last dataset consisted of essays annotated with Big-5 personality traits. Two of these datasets were discussed in detail as part of Chapter 5, while the third (myPersonality Twitter dataset) dataset will be elaborated below. The following section will give a brief summary of the selected datasets and the reasoning behind choosing them.

7.1.1. PAN 2015 Author Profiling Dataset

The first dataset selected was a sample of the PAN 2015 Author Profiling dataset, presented in Section 5.2.3. The dataset was requested directly from the PAN Community [1]. The dataset was chosen as it had proven capable of producing state-of-the-art results for user-level personality prediction Liu et al. (2017). For this Thesis, all Spanish, Dutch and Italian tweets were subtracted, leaving only the English tweets. The resulting dataset consisted of 14,166 tweets from 152 distinct users. Figure 7.1 shows the distribution for

[1] https://zenodo.org/record/3745945#.Ygr5oiyUlhE

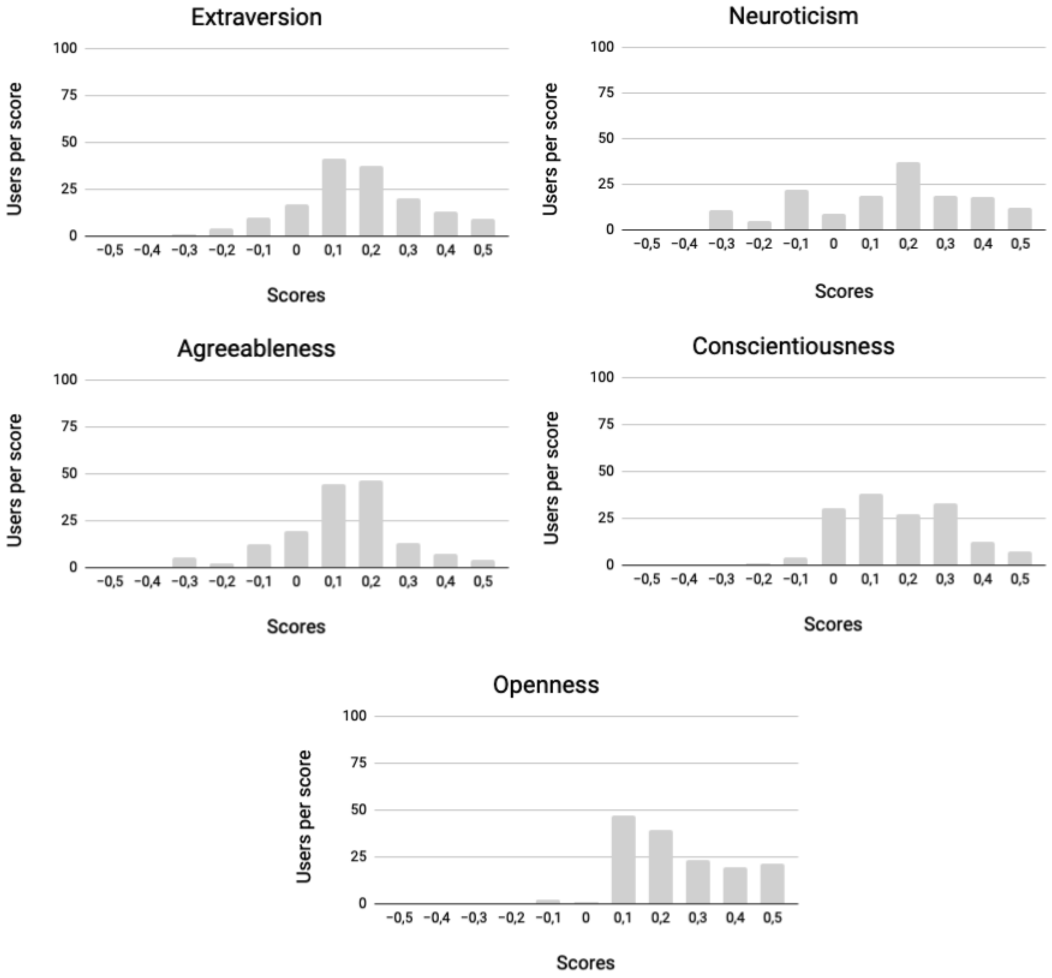


Figure 7.1.: Distribution of Personality Traits for Users - PAN 2015 Author Profiling

each personality trait for each of the users in the sampled dataset. As shown by the figure, extraversion, agreeableness and consciousness has a normal distribution across the users.

7.1.2. myPersonality Twitter Dataset

The second dataset selected was the myPersonality Twitter dataset, a dataset based on the myPersonality Dataset presented in Section 5.2.1. The dataset was obtained from Nordnes and Gran (2019). The original dataset was collected by the myPersonality Facebook Application which allowed users to take a Big-5 personality questionnaire and provide access to their Facebook timeline for research purposes. The application also allowed users owning a Twitter account to provide their Twitter handle. This dataset contains 8,946 tweets collected from the handles of 172 distinct users. Each Twitter user is annotated with the Big-5 personality traits gathered from the initial myPersonality Facebook dataset, with each trait having a score on a continuous scale, ranging from 1 to 5. Similar to the PAN Author Profiling Dataset, the Twitter handle and UserID is replaced by a unique ID, to distinguish the users from one another. As previous studies using the myPersonality Facebook dataset had produced several good performing models for personality prediction and personality classification, as discussed in Chapter 6, the myPersonality Twitter dataset was considered a good fit for this Thesis. Figure 7.2 shows the distribution for each personality trait for each of the users, rounded off to the nearest whole number. As shown by the figure, the distribution is fairly normalised across all traits.

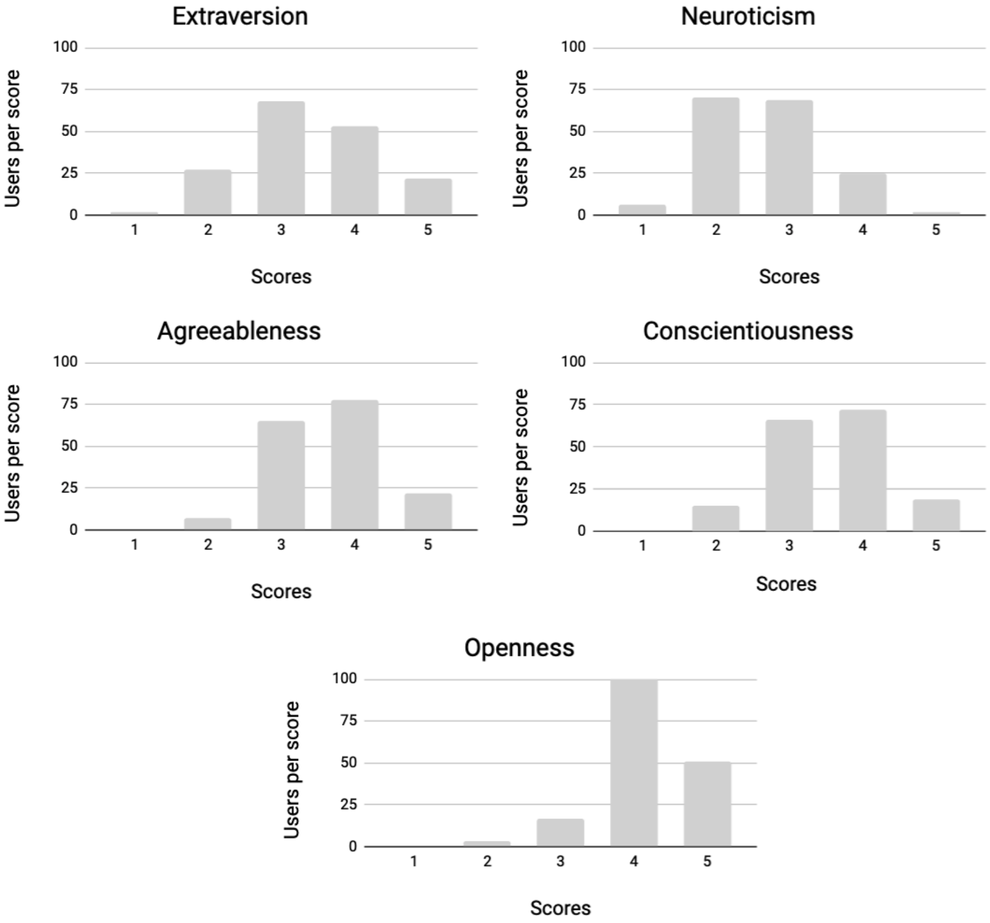


Figure 7.2.: Distribution of Personality Traits for Users - myPersonality

7.1.3. Stream-of-Consciousness Dataset

The third dataset selected was the stream-of-consciousness dataset presented in Section 5.2.4, which consists of essays written by 2,467 psychology students. The dataset was requested through ResearchGate's Forum[2]. The reasoning behind choosing this dataset was that the essays were written in an informal manner, based on whatever came to the students' mind. The informal structure of these essays may therefore share common characteristics with the linguistic style of tweets, making the dataset suitable for training a model for personality prediction from tweets. Below is an excerpt from one of the essays, illustrating the linguistic theme of the essays:

> *Always a problem. My hair is really wet and I should go dry it, but this assignment is what I need to do now. I almost slept through my eight o clock class, but I somehow made it. Ok this show keeps getting cheezier and cheezier oh dear.*

The dataset contains one essay from each student, with an average word length of 653. Each essay is annotated with binary values for each of the Big-5 personality traits, indicating whether the student has the specific personality trait or not.

7.2. Radicalisation Dataset

Section 5.1 presented a selection of datasets used for natural language processing (NLP) tasks within the field of radicalisation. The usage of these datasets varies, from predicting the level of radicalisation influence a user is subjected to, to predicting if a certain user has pro-radical behaviour.

The initial idea of this Thesis was to build a dataset of existing datasets presented in Section 5.1, by using the accounts contained in the datasets as seed accounts for scraping their followers, similar to the approach of Rowe and Saif (2016) and Fernandez et al. (2018). The dataset of Fernandez et al. was requested, but as the owner of the project was on maternity leave during the time of this Master's Thesis, it was not possible to retrieve the data. The dataset of Ferrara et al. was requested, but due to internal and Twitter regulations, the authors were not allowed to share the data. Rowe and Saif were also contacted, but as the authors had both left their positions, the data was not accessible. Lastly, O'Callaghan et al. were contacted, but the inquiry left no reply.

As *How ISIS Uses Twitter*, presented in Section 5.1.3, was the only publicly available dataset, it was chosen for further expansion. Upon exploring the dataset, it was found that 95% of the Twitter accounts contained in this dataset were suspended or inactive as of August 2021. This illustrates one of the challenges when it comes to analysing social media users affiliated with extremism. Several social media platforms

[2] https://www.researchgate.net/post/From_where_can_I_get_a_stream-of-consciousness_dataset

have invested substantial efforts in keeping up with radical content on the web, thereby consistently shutting down these accounts. In a study by Conway et al. (2019), it was found that around 65% of the pro-ISIS accounts and around 20% jihadist accounts were suspended within 70 days of inception. Of the accounts monitored, 30% of the pro-ISIS accounts were shut down within two days, whereas 1% of the jihadist accounts were taken down in the same period.

Based on results from previous studies and the exploration of the *How ISIS Uses Twitter*, it was considered likely that other datasets would have a high degree of suspended accounts. On the basis of these findings it was decided that a new dataset would have to be created for this Thesis. The dataset was built from a set of manually identified Twitter accounts considered to be radicalised. These accounts served as seed accounts, to target users vulnerable to radicalisation. Using the followers of radical users as a base, relevant users were identified using a lexical approach. The approach was similar to that of Fernandez et al. (2018) and Rowe and Saif (2016) who based their annotation on dictionaries containing words associated with extremism. Even though their approaches yielded some promising results, both studies found their annotated dataset of pro-ISIS Twitter users to include non-pro-ISIS users using pro-ISIS rhetoric to reference events associated with extremism or show their disgust towards extremism. Similarly, the hacker community Anonymous experienced challenges when taking down more than 20,000 Twitter accounts allegedly linked to ISIS as a response to the Paris attacks in 2015. The annotation scheme applied to their data resulted in the suspension of the social media accounts of the New York Times, Barack Obama, and the White House [3]. To reduce the risk of including irrelevant users in the dataset, a sample of users retrieved through the lexical approach was assessed manually according to an annotation scheme later described in Section 7.2.3. The following sections will provide information on the constructed dictionary, proposed annotation scheme and data collection approach.

7.2.1. Creating a Dictionary

Identifying users at risk of being radicalised on social media can be a cumbersome and challenging task. As discussed in Chapter 2, radicalisation may unfold differently for every individual and should not be considered as a linear path or a distinct event. Instead, radicalisation is an incremental process in which the path may take several turns leading up to radicalism. Today's radicalisation models are mostly conceptual, and few models have been empirically verified. Still, they share common characteristics which can be conceptualised as a set of indicators of ongoing radicalisation. In order to retrieve relevant accounts for annotation, the indicators were represented in the form of a dictionary. The lexical approach was based on results from previous studies presented in Section 6.2.1. The dictionary was comprised of a selection of terms, from previous studies focusing on jihadism in social media, published radicalisation glossaries and terms identified as

[3] `https://www.bbc.com/news/newsbeat-34919781`

part of studying the language of individuals affiliated with extremism on Twitter. The complete dictionary was comprised of the following:

- **Bodine-Baron et al. (2016)**: a study on ISIS supporters and opposition networks on Twitter. As part of their study they comprised a dictionary of common terms used by ISIS sympathisers.

- **MINDb4ACT** [4]: A glossary of current jihadist terms used by French-speaking jihadists, Salafist-Jihadists, and operational and military jihadists. The dictionary consists of Arabic words and dialects, oral slang forms, and French terms. For this Thesis, only a selection of Arabic words and dialects was included.

- **Digital Jihad**: A glossary developed for the Swedish Justice Department containing terms and propaganda commonly used by ISIS (Cohen and Kaati, 2018).

- **Related Work**: A selection of terms from previous studies on radicalisation verified to have a correlation with jihadism. These studies include Fernandez et al. (2018) Rowe and Saif (2016) and Lara-Cabrera et al. (2017).

- **Identified Terms**: As part of manually identifying radical users on Twitter, a selection of terms affiliated with radicalisation were included in the dictionary.

The terms from previous studies on jihadism and published radicalisation glossaries was selected based on two conditions; **1)** The terms were not ambiguous or open to interpretation, and **2)** There was a certain consensus among the research and glossaries with regards to relevant terms. The terms selected as part of studying the language of users affiliated with extremism on Twitter was based on common language among the users.

Each term was assigned to one of five categories, as presented in Table 7.1. These categories represent common behaviour of individuals at risk of being radicalised (Lara-Cabrera et al., 2017).

[4] `https://mindb4act.eu/wp-content/uploads/2020/04/JihadistVocabulary_MINDb4ACT.pdf` (Accessed 06.11.2021)

Table 7.1.: Lexical Indicators of Radicalisation

Category	**Description**	**Terms**
I1	Frustration and Grievance	*Balec, yomb, intellectual terrorist, polytheist, dirty jew, white christian*
I2	Perception of discrimination	*Sick, hate, discrimination, oppressed, muslim refugee*
I3	Negativity surrounding Western society	*Babtou, west, infidel, hypocrite, al-adu al-baid, oppressors*
I4	Support of Jihadist ideas	*Mujahideen, jihad, martyr, caliphate, daesh, khilafah*
I5	Theological justification for ideas	*Adab, hijrah, kafir, kufr, fāhishah*

I1 represents the feeling frustration and grievance held by an individual regarding their own life situation. As described in Chapter 2, this feeling is considered to be the first step of the radicalisation process.

I2 represents the feeling of discrimination of Muslims in today's society. A common theme among youths on the verge of radicalisation, is a need for justice. These individuals may have a strong feeling of discrimination towards Muslims, resulting in a feeling of unjust (Slootman et al., 2006).

I3 represents a repulsiveness towards Western society and ideas. Activists and recruiters of radical movements tend to nurture the idea of *us-versus-them* thinking, by exploiting personal conflicts and the identity of young Muslims at risk of radicalisation. This idea may in turn provide an ideological template for rebelling towards Western society (Rogers and Neumann, 2007).

I4 represents terms justifying or glorifying jihadist ideas. The terms included in this indicator are terms found to be used by radical individuals and individuals undergoing radicalisation.

I5 represents theological terms directly affiliated with radicals and radicalisation. Theological terms hold a specific role among radicals and during the radicalisation process. Theological references may serve as a catalyst for creating an identity of being part of a Muslim minority and provides an aura of sacredness and authority to the jihadist ideology (Cohen and Kaati, 2018).

While this Thesis focuses on English texts, the dictionary was comprised of both English and Arabic terms. Despite the fact that most Muslims do not hold Arabic as their mother tongue, English jihadists propaganda is filled with Arabic words. Arabic is by many Muslims considered a sacred language and it is not uncommon to mix languages within bilingual minority groups (Cohen and Kaati, 2018). Translating the terms to English was not considered, as this would increase the risk of losing of semantic information.

The Arabic terms contained in the dictionary were all spelled in the Latin alphabet. There is no consensus with regards to Latin spelling of Arabic words. In order to cope with the possibility of neglecting tweets containing relevant content, different variations of spelling were included for certain Arabic terms. The full dictionary is given in Appendix G.

7.2.2. Data Collection

For the purpose of this Thesis, data was collected using Twitter's Academic Research API along with a Python library named Tweepy. The API is available for accepted applicants and is limited to 10 million tweets per month and a certain number of requests per fifteenth minute window. Due to the limitations imposed by the API, the data collection methodology had to be carefully planned. The chosen approach was based on best practices proposed by Parekh et al. (2018). The study reviews previous attempts on building datasets of radical users on social media and proposes a set of guidelines for identifying relevant users. The data collection process is described in the following sections and summarised in Figure 7.3

Identifying Radical Users

In order to generate a set of seed accounts, radical Twitter users first had to be identified. These accounts were manually identified from a list of accounts reported by the CtrlSec[5] and KDK Kill Zone[6] initiatives. These are volunteer organisations targeting extremist users on Twitter and reporting them to Twitter staff through tweets. Users reported over a one-week period, from 22nd October to 28th October, were manually analysed. Users encouraging the use of violence or showing clear support for jihadist ideologies were considered to be radicalised. After the verification, a set of 121 accounts was selected as seeds.

Retrieving Followers

Individuals at risk of mobilising into violent extremism are likely to participate in online sites conveying extremism and seek out relationships with violent extremists [7]. Serving as an indicator of possible mobilisation into radicalisation, all followers of the 121 seed

[5] https://twitter.com/CtrlSec

[6] https://twitter.com/kdktargets

[7] https://www.dni.gov/files/NCTC/documents/news_documents/NCTC-FBI-DHS-HVE-Mobilization-Indicators-Booklet-2019.pdf

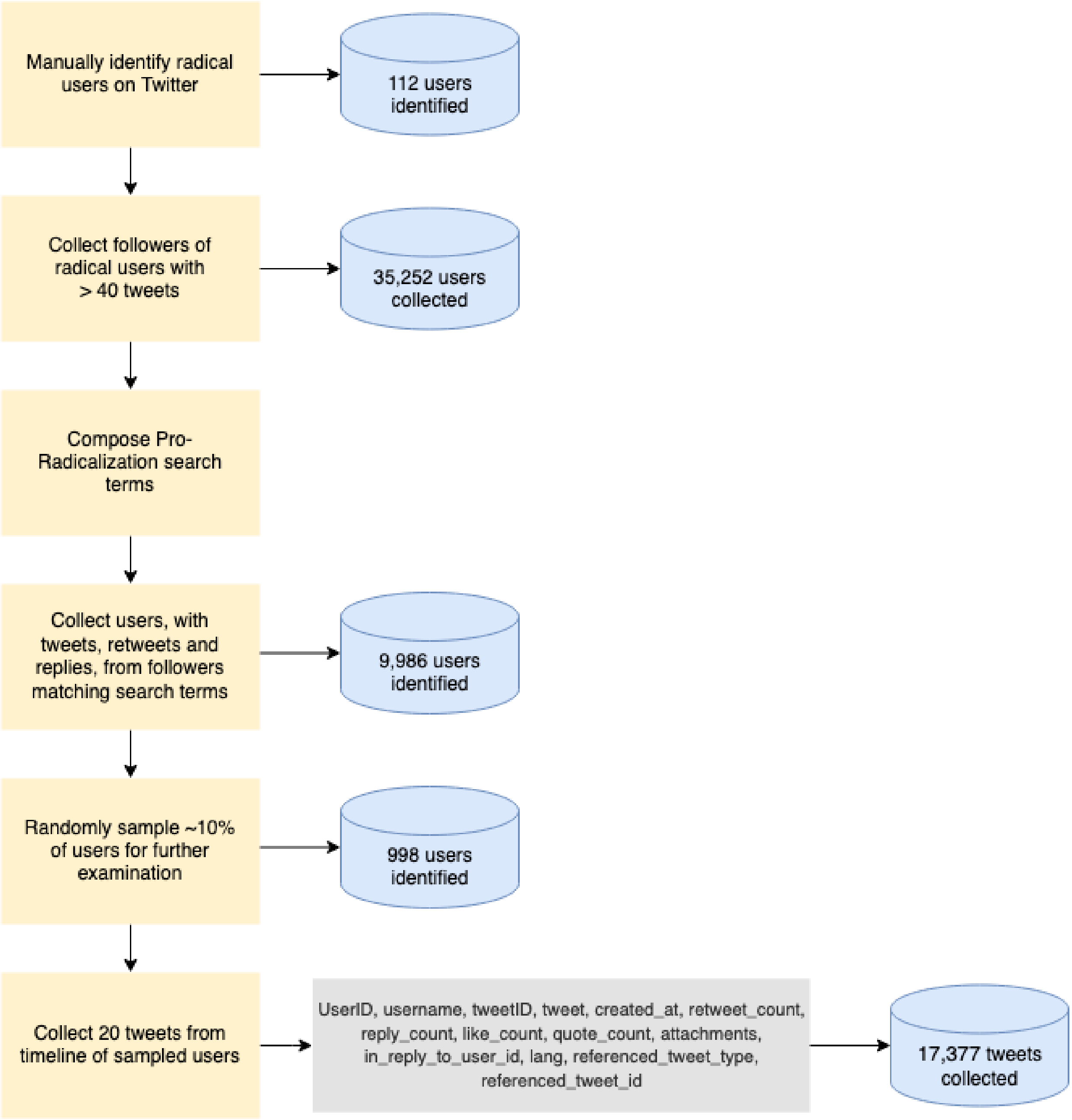

Manually identify radical users on Twitter
112 users identified
Collect followers of radical users with > 40 tweets
35,252 users collected
Compose Pro-Radicalization search terms
Collect users, with tweets, retweets and replies, from followers matching search terms
9,986 users identified
Randomly sample ~10% of users for further examination
998 users identified
Collect 20 tweets from timeline of sampled users
UserID, username, tweetID, tweet, created_at, retweet_count, reply_count, like_count, quote_count, attachments, in_reply_to_user_id, lang, referenced_tweet_type, referenced_tweet_id
17,377 tweets collected

Figure 7.3.: Visualisation of Data Gathering Process

accounts were collected, totaling 72,971 followers. The tweets count of each user was then gathered by querying the API. Followers with 40 or more tweets were included for further analysis, leaving a dataset of 35,252 users.

Targeting Relevant Users

Having built a dataset of relevant users of the seed accounts, the Twitter IDs of each user was then used in a new query. The new query consisted of matching the terms contained in the dictionary against the timeline of each of the users and retrieving the corresponding tweets. For each user, a total of 20 English tweets matching the dictionary were retrieved. For each tweet, the tweet text, tweet ID, username, retweet count, like count, comment count, posting date, referenced tweet ID, reference type, and full retweet were collected. The query resulted in a set of 9,986 users with 97,716 tweets matching the dictionary.

Sampling Relevant Users

From the 9,986 users, 10% of the users were randomly sampled for further exploration. For sampled users with less than 20 tweets matching the dictionary, additional tweets were collected. The resulting dataset consisted of 998 users with 17,377 tweets which were passed on for annotation.

Table 7.2.: Data Collection Statistics

Description	**Number of Users**	**Number of Tweets Collected**
Radical users manually identified on Twitter	121	-
Followers of radical users	72,971	-
Followers with more 40 or more tweets	35,252	-
Users with tweets matching dictionary	9,986	97,716
Randomly sampled users	998	17,377

7.2.3. Annotation

Before creating an annotation scheme, a thorough exploration of existing literature on pathways and indicators of radicalisation had to be conducted. It was important to fully understand the underlying factors contributing to individuals mobilising into violent extremism and how these factors may be observed on social media platforms. It was also important to create a common understanding of how these factors should be represented and how to categorise a user as exhibiting signs of vulnerability towards radicalism. The resulting annotation scheme was heavily influenced by the workings of Rowe and Saif (2016) and Fernandez et al. (2018) who used pro-radical terminology as an indicator for radicalisation and radicalism. While these studies were mainly focused on identifying users already radicalised, this Thesis aim to identify users not yet radicalised. In order to achieve this goal, four additional indicators were proposed, based on common behaviour related to the radicalisation process, resulting in a total of five indicators. The choice of indicators were based on the findings from Chapter 6, the *Mobilization Indicators 2019 Edition*[8], a set of indicators for individuals moving towards violent extremism and a review of conceptual and empirical radicalisation models by Borum (2011a,b) The indicators resemble the ones defined in the dictionary. The following indicators were defined:

1. **Frustration/Grievance:** The users convey feelings of frustration or grievance regarding their own life, society or government.

2. **Negative words/Hateful speech:** The user conveys feelings of hatred, negative sentiment or abusive language towards individuals, communities or government believed to be at fault for their own frustration.

3. **Feelings of discrimination:** The user conveys feelings of discrimination towards Muslim communities.

4. **Negativity towards western society:** The user conveys negative ideas or hatred towards western society.

5. **Pro-Jihadism:** The user endorses, supports or justifies radical ideologies or actions related to jihadism.

Categorising Tweets

Upon defining the indicators, each of the 17,377 tweets were assessed manually and categorised according to the indicators. A total of six categories were defined, five representing the indicators defined in Section 7.2.3 and one named *unrelated*. A tweet could be contained in several of the categories or in none of them. For a tweet not fulfilling any of the indicators, it would be categorised into a unrelated category. An example of categorised tweets are given in Table 7.3.

When categorising according to the indicators, an important distinction was made between endorsement and general sharing of information. It is not uncommon for Twitter users to share violent or disturbing content to the public. Sharing such information does not imply endorsement but may instead be an effort to spread information. Several of the investigated Twitter accounts specified such in their biography. In order for a tweet to be categorised as pro-jihad, there must be a clear intention behind the tweet, that emphasise the endorsement. Table 7.4 gives an example of two tweets that appear similar but are categorised differently. One being a dissemination of information and the other being endorsement.

[8] `https://www.dni.gov/files/NCTC/documents/news_documents/NCTC-FBI-DHS-HVE-Mobilization-Indicators-Booklet-2019.pdf`

Table 7.3.: Example of Tweets Categorised According to Indicators

Example	**Categories**
Afghanistan will not accept dictation from Europeans who bombed us for 20 years and did war crimes. They were droning Afghans on daily basis. Afghanistan is a free country. Foreigners shall not try to dictate and interfere in our affairs.Western culture can't be imposed on Afghans.	1, 4
This rogue slave government of the West licking the shoes of the West whatever it wants! We will not back down from our purpose Go ahead we are ready....MF*** #releaseDrShafeeqAmini	1, 2, 4
Resonant terms that filled the world with its false shouting that the West is highly human, super freedom, and deeply rooted	4
MENTION Hello MENTION , this press conference happened at Press club of India, New Delhi. This hatemonger Narsinghanand abused Prophet of Islam and tried to incite communal hatred in the society. Cc MENTION MENTION #ArrestNarsinghanand #Islamophobia MENTION	1, 3

Table 7.4.: Example of Endorsement Versus Dissemination of Information

Example	Pro-Jihad
Mujahideen captured a group of Afghan students in a small village outside Kabul yesterday. The students were allegedly suspected of preaching polytheism and storing illegal fire arms. Two casualties reported, none of them Mujahideen.	
Mujahideen captured a group of Afghan students in a small village outside Kabul yesterday. The students were allegedly suspected of preaching polytheism and storing illegal fire arms. Two casualties reported, none of them Mujahideen. Let's build this country! #GloryToMujahideen #NoRoomForShirks	✓

Annotating Twitter Users

Once all tweets were categorised, the sampled users were labelled into three classes; (1) Not relevant to radicalisation, (2) Vulnerable to radicalisation, and (3) Already radicalised.

For a user to be labelled as already radicalised, the user would have to publish five or more tweets, retweets, quotes or replies labelled with the *pro-jihadism* category. Users portraying consistent radical behaviour, by five or more tweets, implied radicalism. This annotation scheme was based on the workings of Rowe and Saif (2016) who employed similar measures by defining a cut-off point between radicalism and pre-radicalism. The cut-off point represents the indoctrination phase in the radicalisation process, in which an individual identifies itself with and supports the jihad ideology.

For a user to be labelled as *vulnerable to radicalisation*, they had to fulfill one of the following criteria:

- The user has more than one and less than five tweets, retweets, quotes or replies that falls inside category 5.
- The user has at least one tweet, retweet, quote or reply in at least three of the categories 1, 2, 3, and 4, and less than five in category 5.

Category 5 was considered a high risk indicator, as it represents one of the final and defining steps towards becoming radicalised. For that reason, and similar to that of Rowe and Saif (2016), this indicator was assessed as a stand-alone indicator. Category 1, 2, 3, and 4 were considered weaker indicators of vulnerability towards radicalisation compared to category 5. Being frustrated is by no means a defining indicator of ongoing radicalisation. In order for a user to be labelled as vulnerable to radicalisation, they thus had to have tweets labelled in at least three of the weaker indicators.

For a user to be labelled as *not relevant*, they could not fulfill any of the criteria set for the *already radicalised* and *vulnerable to radicalisation* label. This did not mean that the users are not at risk of being radicalised, but there was not enough evidence from their tweet history to support such a labelling.

After categorising each tweet according to the proposed indicators, 305 of the assessed users were labelled as being vulnerable to radicalisation. This amounts to 30.6% of the randomly sampled users. Considering these statistics for the total number of users matching the dictionary, this amounts to 4.2% of the followers of radical users. In comparison, Rowe and Saif (2016) found 1% of their collected followers to be radical, using only pro-jihadist terms as an indicator. Considering that not all individuals undergoing radicalisation or at risk of radicalisation end up being radicalised, the proportion of users labelled as vulnerable to radicalisation may be considered reasonable.

For the remaining users, 625 (62.6%) were labelled as *not relevant* and 68 (6.8%) as being already radicalised. The resulting annotation is summarised in Table 7.5.

Table 7.5.: Statistics From Annotation

Label	**Number of Users**
Not relevant	625
Vulnerable to radicalisation	305
Already radicalised	68
Total	**998**

While these users were labelled as vulnerable to radicalisation, this Thesis does not claim that these individuals are going to be radicalised, are undergoing radicalisation or are already radicalised. Instead, they exhibit traits that are affiliated with indicators of radicalisation based on radicalisation theory. As discussed in Chapter 2, radicalisation is not a path in which radicalism is the resulting end point. Several individuals undergoing radicalisation leave the path on their way towards radicalism and most individuals never get drawn into radicalisation at all.

In order to ensure sufficient data for personality prediction, additional tweets, quotes and replies were collected for the users labelled as being vulnerable to radicalisation. For each user, up to 100 were collected, including the original data collected from the annotation phase. Retweets were excluded as the personality prediction model makes predictions based on the users own writing style. Upon collecting additional data, it was found that a proportion of the users only contained non-English tweets, quotes and replies, with retweets being in English. Unfortunately, the Twitter API does not offer an easy way to retrieve the tweet count for a specific language, excluding retweets. Due to the lack of non-retweets in English, the resulting dataset was reduced to 278 users. The total number of tweets collected from these users amounted to 15,195, or approximately 55 tweets per user.

Inter-Annotator Agreement

Although the annotation scheme was constructed with clear guidelines for how to interpret each of the six categories and label the tweets, there was still a chance of subjective interpretation of the tweets. In order to test the reliability of the proposed annotation scheme, a random sample of 32 users, with all their tweets, was selected from the 998 users and passed to a second annotator. Each tweet was given binary label by the second annotator for each of the six categories, as per the annotation scheme defined in Section 7.2.3. The total number of tweets amounted to 509, giving a total of 3,054 categories to label. Once labelled, the tweets were assessed and the users annotated according to the annotation scheme defined in Section 7.2.3. The inter-annotator agreement score was then calculated for both the categorisation on a tweet-level and the final annotation on a user-level. The score was calculated using Cohen's kappa and can be found in Table 7.6.

Table 7.6.: Inter-Annotator Agreement for radicalisation$_\mathrm{d}$

	Tweet-Level	**User-Level**
Kappa Score	0.76	0.83

For the 3,054 categories to be labelled in the 509 tweets, the annotators only disagreed upon 207 of them. This yielded a Kappa of 0.76 on a tweet-level, interpreted as *substantial* agreement according to Table 4.1. *not present* or 0 was the most used binary label across the categories and the annotators agreed upon this label 2,444 times. *present* or 1 was the minority label and annotators only agreed upon this label 403 times.

For the final annotation of the 32 users the annotators only disagreed upon 2 labels. This yielded a kappa of 0.83 on a user-level, interpreted as *almost perfect* agreement according to Table 4.1. 23 users were labelled as *not relevant*, 7 were labelled as *vulnerable to radicalisation*, and none were labelled as *already radicalised.*

7.3. Non-Radical Dataset

As part of answering Research Question 3, found in Section 1.2, a second dataset was created, consisting of ordinary Twitter users. The dataset would work as a counterpoise to the dataset of users vulnerable to radicalisation. The intitial idea was to use the *Tweets Targeting ISIS* dataset, described in Section 5.1.4. After exploring the dataset it was found to contain several pro-ISIS users. It was also considered to not be generalised enough to answer Research Question 3. For that reason, the author decided to collect the data manually.

The data was collected by randomly sampling Twitter users with English tweets, by utilising the *Recent Tweets* functionality provided by Twitter. Randomly sampling the users was believed to ensure a diverse representation of users. A total of 259 users were randomly sampled over a period of one day, on January 15th 2022. The number of users was set to equal the number of users found in the final pre-processed dataset of users vulnerable to radicalisation. Up to a 100 tweets was gathered from each of the users, using the Twitter API, resulting in a total of 25,624 tweets.

7.4. Pre-Processing of Personality Datasets

In order to train a personality prediction model, the selected datasets needed to undergo certain pre-processing steps. The resulting personality prediction model would be used on the radicalisation dataset to predict personality traits of users vulnerable to radicalisation. The personality prediction model should be both trained and tested on instances similar that are similar in form. The purpose of the pre-processing step was therefore to ensure a similar structure between the personality datasets and the radicalisation dataset. The personality datasets were already processed to some degree but had to go through further pre-processing steps in order to meet the requirements of a uniform structure. The myPersonality dataset and PAN 2015 Author Profiling dataset were collected both from the same media but had different value ranges for the personality traits. The stream-of-consciousness dataset was based on essays written by individual students and did not hold the same structure as the Twitter datasets. For this reason, different pre-processing steps were required for each of the datasets. Some of the common pre-processing steps applied across the datasets are presented in Table 7.7 while the following sections presents the steps applied to each of the personality datasets.

Table 7.7.: Common Pre-Processing Steps Across Personality Datasets

	Before	**After**
1	? ! , . : ; " - _	
2	1, 2, 3, ... , n	
3	don't	dont
4	HELLO	hello

1. Special symbols were removed from the dataset. Such symbols may be used incorrectly or cause disturbance in the dataset and does not offer much linguistic information

2. Numbers do not offer useful information when training machine learning models and was therefore removed from the dataset.

3. As a step in removing special symbols, left out letters were merged with their associated word. The words *I'm* and *Don't* would be replaced by *im* and *dont.*

4. All text was lowercased to ensure consistency across the datasets and avoid storing several versions of a word in the tokeniser corpus.

7.4.1. myPersonality

The dataset consisted of all tweets from a set of users, with each tweet being represented as a single instance. The first step involved converting all text to lowercase, before replacing all mentions and URLs with *MENTION* and *URL* to ensure a common reference across tweets. As the personality prediction model was to be trained on English language, the next step was to remove all non-English tweets from the dataset. This was done using FastText's built in language detector with the lid.176.ftz library[9]. A total of 364 tweets were removed from the dataset. Once the dataset was cleared of non-English tweets, the symbols listed in Table 7.7 were removed from the tweets. Trailing white spaces were reduced to one. The dataset also contained words with trailing repeating characters which were reduced to two. For instance, would the sentence *"I have a gooooooood feeling about this"* be reduced to *"I have a good feeling about this"*. A selection of common abbreviations or social media slang were replaced by their full form. For instance, would the abbreviated form *dnt* be replaced by *don't.* This step was applied to reduce the chance of out-of-vocabulary words when training the personality prediction model. The list of selected abbreviations can be found in Appendix F.

The last step involved concatenating all tweets from a single user into one coherent string of text. Below is an example of a tweet before and after pre-processing:

Before pre-processing:

[9] `https://fasttext.cc/docs/en/language-identification.html`

> HELLO WORLD! This fall, me n *PROPNAME* are going on our yearly fishing trip to Alaska. You can read more about the trip on our blog http://www.fishingforblokes.com/yearly_fishing_trip.

After pre-processing:

> hello world this fall me and MENTION are going on our yearly fishing trip to alaska you can read more about the trip on our blog URL

7.4.2. PAN 2015 Author Profiling

This dataset was distributed to attendants of the PAN 2015 Author Profiling conference and had already been processed to a certain degree. The pre-processing steps already applied to the dataset differed from the ones proposed in this Thesis and thus the dataset had to be modified to comply with the decided structure and composition. Similar to that of the myPersonality Twitter dataset, all tweets were represented as single instances in the dataset. The tweets were therefore concatenated into one coherent string of text for each of the users. The dataset was split into four parts of English, Spanish, Italian and Dutch, which made it easy to eliminate non-English tweets.

As part of the pre-processing steps already applied to the dataset, mentions were represented as *@username*. To ensure consistency across all datasets, these mentions were replaced by MENTION. URLs were still present in their original form in the dataset and were replaced by the term URL. The dataset also contained references to images and videos. These were represented as *Photo:* and *[pic]:* for images and *Video:* for videos. The references were replaced by *PHOTO* and *VIDEO*.
Several of the hashtags contained in the data were made up of multiple English words. These words were split into separate words by using the *wordsegment* library[10]. For instance, would the hashtag *#readytoparty* be replaced by *ready to party*.

Similar to that of the *myPersonality Twitter Dataset*, trailing white spaces was reduced to one, trailing repeated characters were reduced to two, and abbreviations replaced by the full form.

A number of retweets were still present in the dataset. The personality prediction model of this Thesis focuses on text written by the users themselves. For that reason, all retweets were removed, reducing the dataset by 315 tweets.

The tweet instances of the dataset were annotated with personality traits on a user-level, with a score ranging from -0.5 to 0.5. In order to comply with the myPersonality Twitter dataset, this scale was converted to a scale ranging from 1 to 5 for each trait.

[10] `https://pypi.org/project/wordsegment/`

Finally, last step of the pre-processing involved the five steps presented in Table 7.7. The example below shows a tweet before and after pre-processing:

Before pre-processing:

> Video: @username, please watch this video, you would benefit from doing so! I would also recommend that you read this article from http://www.article.com/on_something before making such claims... #ignorantfool

After pre-processing:

> VIDEO MENTION please watch this video you would benefit from doing so i would also recommend that you read this article from URL before making such claims ignorant fool

7.4.3. Stream-of-Consciousness

The stream-of-consciousness dataset consisted of single essays from individual students and did not require any concatenation. The dataset was pre-processed in a similar fashion to that of the PAN 2015 Author Profiling and myPersonality dataset, by performing each of the steps presented in Table 7.7, except having to remove hashtags. In contrast to the Twitter datasets, the personality traits of the stream-of-consciousness dataset were labelled with binary values. In order to train a regression model for personality prediction, these values needed to be converted from binary classification labels to numerical values. The conversion was based on the approach and findings presented by Nordnes and Gran (2019). Their approach took advantage of numerical and binary values contained in a collection of Twitter datasets. For each personality trait, and average numerical value was calculated for the binary values, for each personality trait. Their approach was tested using Global Vectors with 200 dimensions and Gaussian Process, yielding promising results for the majority of the traits. Once the average numerical values were calculated from the Twitter datasets, the binary values were then supplemented by numerical values according to their calculation. The calculated values for the stream-of-consciousness dataset were based on a combination of the myPersonality Twitter dataset and the PAN 2015 Author Profiling Dataset. An example of the conversion is shown in Table 7.8.

Table 7.8.: Numerical Value Representation for Binary Label

cEXT	**cAGR**	**cNEU**	**cCON**	**cOPN**	**sEXT**	**sAGR**	**sNEU**	**sCON**	**sOPN**
y	y	y	y	y	4.00	3.93	3.83	4.00	4.17
n	n	n	n	n	2.74	2.92	2.24	2.92	3.29

7.4.4. Accumulated Pre-Processed Personality Datasets

Once each personality dataset had been pre-processed separately to the same format, they were concatenated into one coherent dataset, referred to as $personality_d$. This dataset served as a basis for training the personality prediction model. Table 7.9 shows the statistics for each dataset separately and the accumulated dataset *personality*$_d$.

Table 7.9.: Statistics for Personality Datasets

Dataset	**Number of Users**	**Shortest Text**	**Longest Text**	**Average Text**
myPersonality Twitter Dataset	172	8 words	5,123 words	733 words
PAN 2015 Author Profiling Dataset	152	353 words	1,983 words	1,194 words
Stream-of-consciousness Dataset	2,467	35 words	1,120 words	653 words
Personality$_d$	2,791	8 words	5,123 words	687 words

7.5. Pre-Processing of Radicalisation Dataset

The collected and annotated dataset of users vulnerable to radicalisation, hereby referred to as *radicalisation*$_d$, consisted of 278 users with an average of 55 tweets per user. Compared to the personality datasets, this dataset had no prior pre-processing steps applied to it. In order to comply with the defined format, several steps had to be taken. As opposed to the personality datasets, *radicalisation*$_d$ contained several emojis within each tweet. Emojis can reveal information about emotions expressed in text that in turn can help improve upon contextual features. However, as the personality datasets were already stripped of emojis, it was decided to keep a uniform format across all datasets. The first step therefore included removing these emojis from the dataset. For this, the PyPi emoji library was used [11]. Once removed, each tweet was converted to lowercase, before replacing mentions and URLs by *MENTION* and *URL*. As mentioned in Section 7.2.3, retweets were excluded from *radicalisation*$_d$, and did not have to be accounted for as part of the pre-processing.

As part of the collection process, the API query was set to only scrape English tweets. However, several of the tweets collected contained single terms or phrases written in non-Latin letters. Translating these terms and phrases to English may result in loss of semantic information. The next pre-processing step was therefore to remove all non-Latin letters. To ensure that the dataset only contained English tweets, the tweets were checked using the same approach as for myPersonality. Once stripped for any non-English tweets, all special characters and numbers were removed, as per Table 7.7. The final steps involved replacing abbreviated words, removing trailing characters

[11] https://pypi.org/project/emoji/

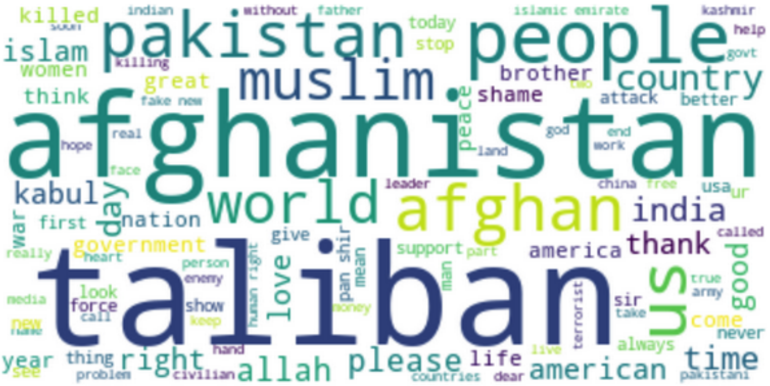


Figure 7.4.: Most Common Terms for *radicalisation*$_d$

and white spaces, and segmenting hashtags, similar to the PAN 2015 Author Profiling pre-processing, before concatenating the tweets of each used.

Several tweets were removed due to a high degree of non-English terms. For this reason, 19 users were removed from the dataset due to having a word count of 40 or less terms, after concatenation. The resulting dataset consisted of 259 users.

As part of exploring the pre-processed data, a word cloud was created using the PyPi wordcloud library [12]. The wordcloud was meant to give insight into the most common terms among the users and used for comparison with the composed dictionary. Stopwords were removed prior to generating the word cloud using the built in funcionality of wordcloud. Additional stopwords such as 'URL' and 'MENTION' were appended to the stopwords list to avoid these terms from appearing in the wordcloud. The wordcloud can be seen in Figure 7.4 and is further discussed in Section 10.1.

7.6. Pre-Processing of Non-Radical Dataset

Similar to that of *radicalisation*$_d$, the dataset of non-radicals, hereby referred to as *non-radical*$_d$, had to undergo several pre-processing steps. Almost an exact copy of the pre-processing algorithm used on *radicalisation*$_d$ was used for *non-radical*$_d$. The only difference was that no non-Latin letters were found to be precent in *non-radical*$_d$ and thus did not have to be removed.

Once pre-processed, the dataset was explored by generating a word cloud for the most common terms among the users, in a similar way to that of *radicalisation*$_d$. The word cloud was meant to serve as a comparison for the terms used between *radicalisation*$_d$ and *non-radical*$_d$. The word cloud can be seen in Figure 7.5.

[12] https://pypi.org/project/wordcloud/

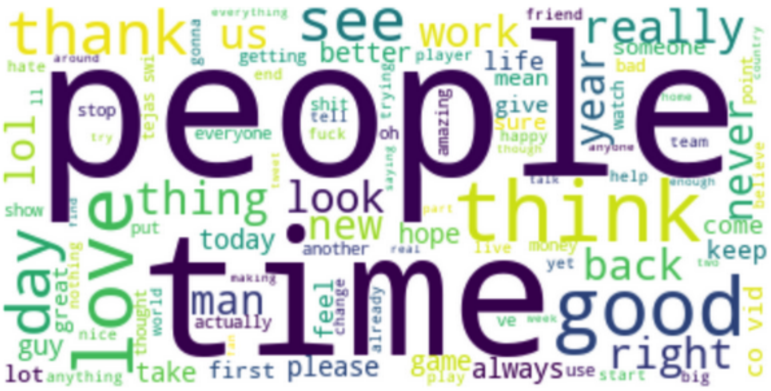


Figure 7.5.: Most Common Terms for *non-radical*$_d$

8. Architecture

The following chapter will describe the architecture of the proposed personality prediction models used in the experiments of this Thesis. The first section presents the feature extraction policies applied for training and testing the models. The second section presents the architecture of each of the implemented models. This chapter gives an overview of the architectures, while Chapter 9 elaborates further on the implementation and selection of hyperparameters.

8.1. Feature Extraction

The performance of a model is only as good as the data that it is trained on. The choice of features for representing a text may therefore have a tremendous effect on the performance of the model. The number of selected features also impacts the performance. A greater number of features may offer more information for the model to train on and make its predictions but may also cause the model to overfit the training data or call for more computational power. For this Thesis, a selection of feature extraction approaches was tested with different models, to derive an architecture best suited for predicting personality traits from linguistic queues. The choice of methodology was based on the reviewed literature and tested separately against a baseline personality predictor created specifically for this Thesis. The emphasis was put on word embeddings generated by pre-trained transformer-based models as they have proven to generate highly contextualised features, capable of producing state-of-the-art performance across multiple Natural Language Processing (NLP) tasks. The below sections provide an overview of the selected feature extraction methodologies and the rationale behind choosing these features. The theoretical framework for understanding the methodologies is described in Chapter 3.

8.1.1. LIWC

As part of building a baseline prediction model for this Thesis, the Linguistic Inquiry and Word Count 2015 (LIWC2015) software was used for feature extraction. LIWC, explained in Section 4.1.2,is an effective and easy-to-use tool for extracting features, such as positive and negative emotions, which in turn can provide useful insight into the personality traits of individuals. LIWC outputs a total of 93 features, covering the emotional, cognitive and structural components present in a written text. LIWC features for classification and regression has been used in previous studies to create models and baseline predictors for personality, among them being Tandera et al. (2017), Xue et al.

(2018), and Farnadi et al. (2021). Due to the effectiveness and previous results of LIWC for personality prediction, it was considered a good fit for creating a baseline predictor.

8.1.2. Term Frequency-Inverse Document Frequency

TF-IDF, explained in Section 4.1.2, uses statistical measures to evaluate the importance of words in a document across the corpus and extract relevant information. As found in the reviewed literature and summarised in Table 6.1.3, TF-IDF is a commonly used feature extraction methodology for training personality prediction models, yielding promising results. By identifying relevant terms across a collection of documents, these features could provide useful information to a personality prediction model for identifying correlations between relevant terms and personality traits. TF-IDF is easy to use and a computationally efficient methodology and was therefore considered a good stepping stone for training a model and comparing the performance against the baseline predictor, trained and tested on LIWC features. The documents were stripped of English stopwords, before being tokenised into unigrams of words. The resulting tokenised documents served as a basis for the TF-IDF vectorisation scheme.

8.1.3. Word Embeddings

One of the drawbacks of TF-IDF and LIWC is that these features do not consider the similarity between words. Word Embeddings allow for creating a distributed representation based on the usage of words, meaning that words that appear together result in having similar representations. Word Embeddings have contributed to vast advances in the field of NLP by their ability to capture similarity. Among them being Arnoux et al. (2017), who used GloVe word embedding in their study of personality prediction from tweets, outperforming the current state-of-the-art with eight times fewer input data. On the basis of the results from previous studies utilising word embeddings, the methodology was chosen for this Thesis.

While GloVe has proven to generate useful features for personality prediction, it assigns the same value to a word, regardless of the context it appears in. In order to overcome these limitations, two pre-trained transformer-based models where chosen for generating word embeddings; namely DistilBert (Sanh et al., 2020) and ALBERT (Lan et al., 2020), described in Section 3.1.9 and Section 3.1.10. The architecture of these models is based on the original BERT architecture (Devlin et al., 2019), but are less computationally heavy while yielding results close to, and sometimes better, than some of the BERT models.

It is possible to train the models manually on a dataset but considering the sparsity of the *personality*$_d$ dataset, it was decided to implement the pre-trained models, more specifically the *albert-base-v2* and *distilbert-base-uncased.*

Each document in $personality_d$ was tokenised sequentially using the built-in encoder algorithms of the respective models. DistilBert uses WordPiece for tokenisation, explained in Section 4.1.2, while ALBERT uses SentencePiece, explained in Section 4.1.2. Both models require the use of special tokens in order to process the documents. A [CLS] token is used to mark the start of the document, while a [SEP] token is used to mark the separation between sequences. The [SEP] token is used at the end of each sequence, even when only one sequence is passed to the encoder. The documents were passed to the tokeniser, adding the special classifier token [CLS] at the start of each document and [SEP] at the end of a document, and converting the tokens to indexes. Out-of-vocabulary words were handled by the tokenisers by splitting the words into subwords contained in the WordPiece and SentencePiece vocabulary.

Both models accept a max position embedding of 512 tokens by default, while the average word length of $personality_d$ was 687 words. Documents surpassing the limit of 512 tokens were truncated, keeping the first 512 tokens of the document. Documents with less than 512 tokens were padded with a special pad-token at the end of the document to create equal length input for the model. An illustration of the process can be seen in Figure 8.1.

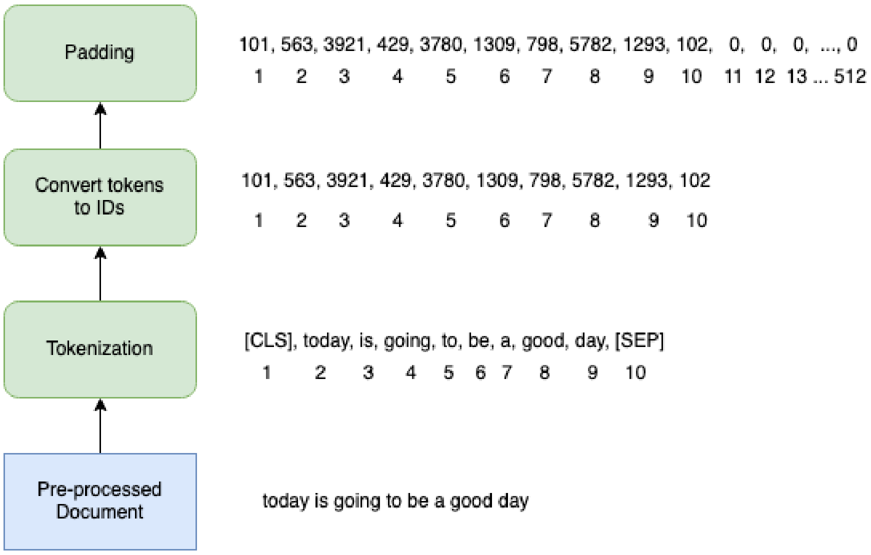


Figure 8.1.: Tokenisation Pipeline for ALBERT and DistilBert

The output from the tokenisers were then fed into the main component of the architecture, which consists of a series of fully connected transformer encoder blocks. ALBERT is similar to that of BERT base and consists of 12 layered encoder blocks. DistilBert is a distilled version of BERT, consisting of only 6 layered encoder blocks.

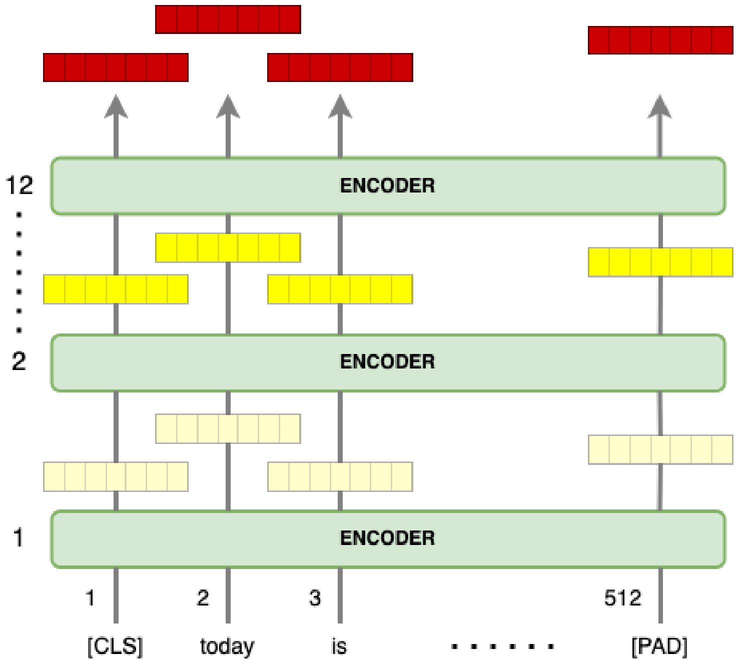


Figure 8.2.: Word Embedding Pipeline for ALBERT. Adapted from Alammar (2018), with permission from Jay Alammar.

The tokenised input is fed to the first encoder layer along with segment ids for the input, one for each token. The segment ideas are set to 1 if only one sequence is used. The output of the encoder block is a word embedding of dimension 768 which is then passed as input to the next encoder in the layer. The final layer outputs a word embedding vector for each of the 512 tokens, yielding an embedding matrix of dimension 512x768. An illustration of the ALBERT architecture can be seen in Figure 8.2.

In order to create a feature vector for the prediction models to train on, a sentence embedding was generated using the word embeddings of the document. The sentence embedding was taken as the average of the 512 word embedding vectors. An illustration of the process can be seen from Figure 8.3. For this Thesis, sentence embeddings were extracted from layer 2, 11 and 12 of ALBERT and layer 2, 5 and 6 for DistilBert. It is possible to generate sentence embeddings from either of the layers of ALBERT and DistilBert. Typically, as embeddings move deeper into the network, they pick up more contextual information. However, as they move deeper, they are also more likely to pick up information related to the initial pre-training tasks of the models. Which layer produces the best embedding may depend very much on the task at hand. For the reason of testing the performance of the embeddings, sentence embeddings were selected from the first and last layers of the architecture and tested for personality prediction.

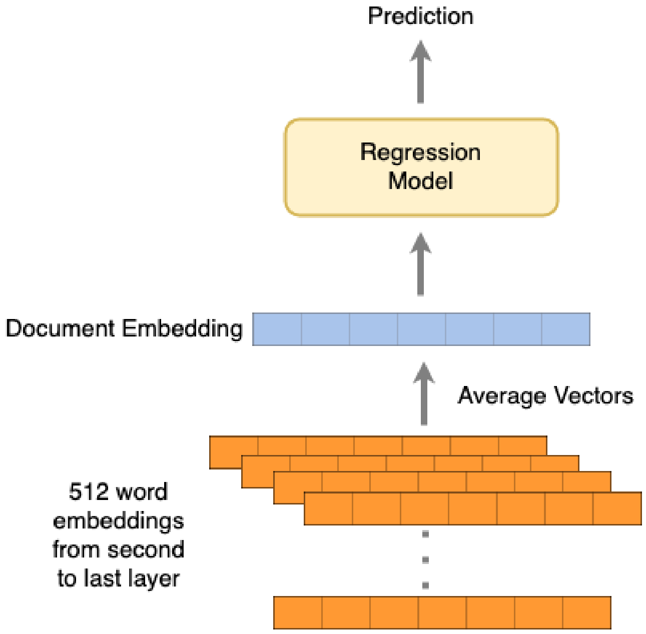


Figure 8.3.: Sentence Embedding Pipeline for ALBERT

8.2. Personality Prediction Model

For the purpose of this study, a total of four machine learning algorithms were explored. The Support Vector Regression, Gaussian Process and Multi-Layer Perceptron were selected based on related work reviewed as part of Chapter 6. The Ensemble learning architecture was chosen by the author in an effort to explore alternatives for improving upon the performance of the models. Ensemble models are known for better capturing non-linear relationships which in turn can result in better performance, and the architecture was thus considered a good fit for the task at hand. Each of the architectures are described in detail in Section 3.1.

The architectures were trained and tested using the different feature groups described in Section 8.1. An illustration of the pipeline can be seen in Figure 8.4. The models were tested with multiple configurations of the architecture by running a Grid Search across a set of parameters. The implementation is discussed further in Section 9.2

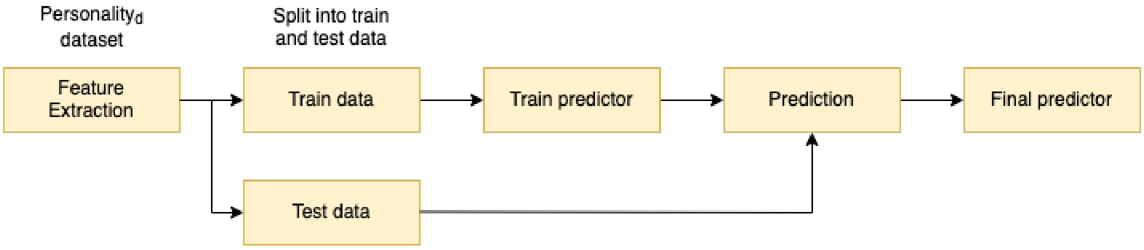


Figure 8.4.: Pipeline for Personality Models

8.2.1. Support Vector Regression Architecture

In order to build and test the performance of different model architectures, a baseline model was created to serve as a reference point when evaluating the models. The Support Vector Regression (SVR) model, explained in Section 3.1.3, was chosen, as it is a popular machine learning architecture, used in both classification and prediction tasks of personality traits. As can be seen from Table 6.1.3, the model has proven to gain good results in previous studies on personality detection and prediction. LIWC features in combination with SVM (SVR) has been used in previous studies to create models and baseline predictors for personality, among them being Tandera et al. (2017), Xue et al. (2018), and Farnadi et al. (2021). Due to the effectiveness and previous results of combining LIWC with SVR, it was considered as a good fit for creating a baseline predictor. The baseline SVR was implemented using a linear kernel due to the computational efficiency of the model.

The SVR architecture was also used for training models for comparison with the baseline predictor. These models were trained on each of the extracted feature groups separately. Three variations of the architecture were tested, with the variations being in the use of kernel, namely *RBF*, *sigmoid*, and *linear*. Each of the kernels were tested on the same parameters, by running a Grid Search across a selection of parameters.

8.2.2. Gaussian Process Architecture

The Gaussian Processes (GP) architecture, explained in section 3.1.6, was chosen on the basis of previous results obtained by Golbeck et al. (2011a) and Arnoux et al. (2017). Golbeck et al. (2011a) used GP for predicting personality traits of Twitter users with predictions being within 11-18% of the actual value. Arnoux et al. (2017) study obtained a new state-of-the-art with eight times less data, using word embeddings as features. Due to the resemblance between these studies and this Thesis, GP was considered to be an interesting candidate.

Two variations of the architecture were tested; one as the sum of the *RBF* and *white-kernel*, and another as the sum of the *dot-product* and *white-kernel*. White-kernels are used as part of a sum-kernel and particularly useful when it comes to explaining noise in data. Both kernels were tested on the same parameters by running a Grid Search across a selection of parameters.

8.2.3. Multilayer Perceptron Architecture

The Multilayer Perceptron (MLP) architecture, explained as part of Section 3.1.8, was chosen for two reasons; **1)** The previous studies on personality prediction by Lima and de Castro (2014) and Xue et al. (2018) had generated interesting results, and **2)** MLPs are generally good at training on high dimensional data. Similar to that of Xue et al. (2018), the model was trained on deep semantic features, using the embeddings generated by ALBERT and DistilBert. Multiple variations of the architecture, among them the number of layers, was tested by running a Grid Search across a selection of parameters.

8.2.4. Adaptive Boosting Architecture

Adaptive Boosting, or AdaBoost is a type of ensemble learning, explained in Section 3.1.11. The Adaptive Boosting Architecture was chosen for two reasons; **1)** its ability to learn from past prediction errors and correct its behaviour while training, and **2)** its ability to cope with and curb over-fitting. The latter was considered important as the trait scores from the essay dataset, contained in $personality_d$, were given as the average float score of each binary value from the myPersonality Twitter dataset and PAN2015 Author Profiling Dataset. The architecture was implemented with Regression Decision Trees as the base estimator, using boosting to generate a series of weak learners, or "stumps", to derive one strong learner. Multiple variations of the architecture was tested by running a Grid Search across a selection of parameters.

9. Experiments and Results

The following chapter will cover the conducted experiments and their results. The first part off the chapter starts off by presenting the experimental plan and which research question they are aimed at answering. The second part includes the experimental setup and covers the technologies used and their implementation. The last part will present the results of the conducted experiments, which will be further discussed in Chapter 10. The experiments are oriented around Research Question 3 and 4, and the overall goal of this Thesis. Research Question 1 and 2 are covered in Chapter 10 as part of evaluating the proposed annotation scheme.

9.1. Experimental Plan

In order to ensure a structured experimental process, an experimental plan was developed. The plan consisted of three parts, aimed at answering the final two research questions of this Thesis. Each of the parts are carried out as a series of experiments and build upon each other. The first part describes the development of a regression model for personality prediction. The model is trained and tested on the $personality_d$ dataset using the features presented in Chapter 8. The second part presents the experiment for predicting personality traits of individuals at risk of radicalisation, using a selection of models derived from Experiment 1. The experiment is conducted using the manually annotated $radicalisation_d$ dataset and compared with results from predictions on $non\text{-}radical_d$. The third and final part presents the experiment for testing the correlation between indicators vulnerability towards of radicalisation and the traits predicted from $radicalisation_d$.

9.1.1. Experiment 1

The first experiment carried out is a preparation for experiment 2 and aims create a personality prediction model that can be used for answering Research Question 3, presented in Section 1.2. A model is trained with the same architecture for each of the Big5 personality traits, and the best models is passed on for experiment 2. All models are trained and tested on the same dataset, namely the $personality_d$ dataset, to derive the best performing model for personality prediction. The experiment is carried out using all architectures presented in Chapter 8, with a combination of features. To measure the performance of the individual models, a baseline predictor is created using SVR with LIWC features, as presented in Section 8.2.1. The baseline predictor is implemented with a *linear* kernel, with the performance tested against *RBF* and *sigmoid* kernels. The kernels are described in detail in Section 3.1.7. Due to continuous state-of-the-art results for transformer-based word embeddings, the emphasis is put on these features. Word embeddings are therefore the only features tested on all models, while LIWC and TF-IFD are only used in combination with SVR. A summary of the features and models to be tested are given in Table 9.1:

Table 9.1.: Overview of Features Tested on Models

Model	**LIWC**	**TF-IDF**	**DistilBert Embeddings**	**ALBERT Embeddings**
SVR	x	x	x	x
MLP			x	x
Gaussian Process			x	x
Adaptive Boosting			x	x

Fine-tuning and Optimisation

To create a model capable of performing personality prediction with reasonable accuracy, each model is fine-tuned and optimised using grid search across a selection of parameters. The grid search is performed with a 5-fold cross-validation with Pearson Correlation Coefficient as the score metric, using $personality_d$ for training and validation. The train-test split is made up of a 90/10 split of the original $personality_d$ dataset, and the train set split further into train and validation sets of approximately 90/10. The average of the five folds is returned along with the parameter configuration and used to assess the performance of each model. The parameters are further elaborated in Section 9.2.

In order to find which embedding layer of ALBERT and DistilBert capture the most contextual information, sentence embeddings from layer 2, 11 and 12 are gathered and tested. Each of these embeddings are used for training the models on each trait respectively. In an effort to find the best feature representation for the data to train on, the word embeddings of each layer are extracted both with and without stopwords being removed. What is considered to be stopwords may depend upon the domain in which the model is being trained on. This results in a total of six embedding variations for ALBERT and DistilBert.

Testing

For testing the performance of each model, the test set from the train-validation-test split of the $personality_d$ dataset is used. As the models are optimised using a selection of linear and non-linear kernels, the coefficient of determination is discarded and replaced with Pearson Correlation Coefficient as score metric during the cross-validation. The performance of each model returned from the grid search is further validated by calculating the Pearson Correlation Coefficient between the predicted trait score and the true score of the test dataset.

Model selection

When all models have been tested, each model is assessed according to the score metric and the best performing model will be chosen for Experiment 2. As one model of each architecture is trained for each trait, the average performance is compared between the models. For the case of there being no best performing model across all traits, the best performing model for each trait will be selected.

9.1.2. Experiment 2

Experiment 2 aims to answer Research Question 3 and will explore the predicted personality traits of the *radicalisation*$_d$ dataset and *non-radical*$_d$ dataset. For prediction, the best performing model of each trait is selected from Experiment 1. The datasets are then passed through the same tokenisation and vector-representation scheme for feature extraction, before being passed to the model for prediction. The mean trait score for each dataset is calculated along with the standard deviation for comparison.

9.1.3. Experiment 3

Experiment 3 aims to answer Research Question 4 by exploring the relationship between the defined indicators of ongoing radicalisation and the predicted personality trait scores from the *radicalisation*$_d$ dataset. In order to explore the relationships, the Pearson Correlation Coefficient is used. The aggregated indicator score for each Twitter user from the annotation process, described in Section 7.2.3, is used to assert any correlation with the predicted traits from Experiment 2. Each indicator and trait is assessed separately, and the strength of the correlation is used as a metric to answer Research Question 4.

9.2. Experimental Setup

This section will cover the experimental setup of each experiment with the aim of making them reproducible. Each experiment is listed sequentially and include information on the implementation, configuration, libraries and parameters.

9.2.1. Personality prediction model

In order to train a personality prediction model the features described in Section 8.1, are extracted. To find the optimal combination of parameters, a grid search was performed on each of the models using the GridSearchCV library from Scikit Learn [1]. The *GridSearchCV()* is implemented with the following configurations:

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

- refit=True: to refit the estimator using the best found parameters
- verbose=3: returns the score and parameters for each fold. Used to keep track of the training process.
- scorer=custom_scorer: custom scoring metric appended to the model using the *make_scorer* function from *sklearn.metrics*[2]. Custom metric is set to Pearson Correlation Coefficient.

LIWC

LIWC was explained in section 4.1.2. To extract the LIWC features, each document from the $personality_d$ dataset was passed through the LIWC2015 text analyser tool. The analyser returned a total of 93 features for each of the documents, represented as numerical values. All of the features were included in the training dataset and used for the baseline predictor.

TF-IDF

TF-IDF was introduces as part of section 4.1.2. To extract the TF-IDF features, the sklearn.feature_extraction.text.TfidfVectorizer library[3] from Scikit Learn is used. Using TfidfVectorizer, each document was tokenised into n-grams of words and stopwords removed, before being converted to the TF-IDF vector representation. The TfidfVectorizer was initialised with the following parameters:

- ngram_range=(1,1)
- stopwords='english'
- max_features = 9000

Word Embeddings

Word Embeddings are extracted from the pre-trained ALBERT and DistilBert models, both explained in Section 3.1.9 and Section 3.1.10. The transformer-based models are implemented using the *transformers* library from Huggingface [4]. Due to the sparsity of the dataset, the models were not trained on the $personality_d$ dataset.

Using the *encoder* function of the *AlbertTokenizer* and *DistilBertTokenizer*, each document was tokenised with special tokens [CLS] and [SEP], and converted to vocabulary indices. Tokenised documents with less than 512 tokens were padded with

[2] https://scikit-learn.org/stable/modules/model_evaluation.html#scoring

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[4] https://huggingface.co/docs/transformers/index

zeroes to reach a length of 512. For tokenised documents with more than 512 tokens, the documents were truncated to reach a length of 512, removing the last tokens.

Each of the models were implemented with default parameter settings, except for the *output_hidden_states* parameter which was set to *True* in order to retrieve the word embeddings.

As discussed in Section 8.1.3, three sentence embeddings were created from layer 2, 11 and 12 of ALBERT by taking the average of the word embeddings. Similarly, three sentence embeddings were created from layer 2, 5 and 6 of DistilBert using the same approach. Sentence embeddings were created with and without stopwords, yielding a total of six sentence embeddings from each of the models. Stopwords were removed using the *gensim 4.1.2* library from PyPi [5].

SVR Implementation

The SVR was implemented using the *sklearn.svm.SVR* from Scikit Learn [6]. The model was trained and tested on all features separately, for each of the traits. In order to find the best configuration of parameters, the model was run with a grid search. The grid search is initialised with the following parameters:

- C: [0.1, 0.5, 1, 5, 10, 15, 20]
- kernel: ['rbf', 'linear', 'sigmoid']
- epsilon: [0.1, 0.2, 0.5, 0.8]

The baseline model is only implemented with a linear kernel, using LIWC features. LIWC features are also tested with the *RBF* and *sigmoid* kernel and run through a grid search, but only for comparison with the established baseline.

MLP Implementation

The MLP is implemented using the *sklearn.neural_network.MLPRegressor* library from Scikit Learn[7]. Due to the model's ability model to deal with high dimensional data, the models are trained and tested on each of the sentence embeddings derived from the transformer-based models. Each model is run through a grid search to derive the optimal set of parameters for each trait. The grid search is initialised with the following parameters:

- hidden_layer_sizes: [(50,), (100,), (200,), (500,)]

[5] https://pypi.org/project/gensim/

[6] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

- activation: ['logistic', 'tanh', 'relu']
- learning_rate_init: [0.001, 0.005, 0.01, 0.05, 0.1]

Gaussian Process Implementation

The Gaussian Process is implemented using the *sklearn.gaussian_process.GaussianProcessRegressor* library from Scikit Learn [8]. The model is trained and tested on each of the sentence embeddings derived from the transformer-based models. Each model is run through a grid search to find the optimal kernel. All other parameters are set to default. The tested kernels are:

- DotProduct() + WhiteKernel()
- RBF(length_scale=1.0) + WhiteKernel()

Adaptive Boosting Implementation

The adaptive boosting is implemented using the *sklearn.ensemble.AdaBoostRegressor* library from Scikit Learn [9]. The *base_estimator* is set to default, with a boosted ensemble using *DecisionTreeRegressor* [10] with a *max_depth* of 3. The model is trained and tested on each of the sentence embeddings generated from the transformer-based models. Each model is run through a grid search to derive the optimal set of parameters for each trait. The grid search is initialised with the following parameters:

- n_estimators: [50, 100, 150, 200, 300]
- learning_rate: [0.001, 0.01, 0.1, 0.5, 1]
- loss: ['linear', 'square']

Environmental Resources

The feature extraction and optimisation of the models was conducted using IDUN (Själander et al., 2021), a high-performance computing cluster hosted by NTNU. The cluster enables the use of the NVIDIA Tesla V100 and P100 GPUs. Models trained on LIWC and TF-IDF features took approximately 3 hours to run, while models trained on high dimensional sentence embeddings from the transformer-based models took up to 8 hours. The AdaBoost models took the longest, with training time approaching 12 hours for certain models.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html

[9] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html

[10] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor

9.3. Experimental Results

This section introduces the results of the experiments conducted as part of this Thesis. Each experiment is presented separately, in incremental order. First, the results of the best performing models of each feature-model-combination is presented for each of the traits. The best performing models are those returned from the performed grid search. Secondly, the results from the personality traits prediction for *radicalisation*$_d$ and *non-radical*$_d$ are presented. Lastly, the estimated correlation between indicators of vulnerability towards radicalisation and the predicted traits of *personality*$_d$ are presented.

9.3.1. Experiment 1

For Experiment 1, the models were trained on multiple runs with different features and using a grid search with a set of predefined parameters. The results of the best performing models of each trait is summarised in Table 9.13. All results were measured using the Pearson Correlation coefficient. SVR was the only model trained and tested with all features. The Gaussian Process, Adaboost and MLP models were trained and tested using sentence embeddings generated from the transformer-based models.

Establishing a Personality Prediction Baseline

The results of the baseline predictor can be seen in Table 9.2. The baseline predictor was trained on all 93 features produced by the LIWC text analyser. The results are somewhat interesting. While the model performs poorly for extraversion and agreeableness, with close to no correlation, the model performs surprisingly well for openness. This may suggest that the LIWC features are able to pick up information about a person's openness, as opposed to the remaining traits. For neuroticism and conscientiousness, the model yielded results showing a weak positive correlation. The correlation for these traits were still considered to be low with regards to predicting one's personality.

Table 9.2.: Score for Baseline SVR with LIWC Features

Extraversion	**Agreeableness**	**Neuroticism**	**Openness**	**Conscientiousness**
-0.08	0.04	0.12	0.27	0.11

The best trait estimators returned from the grid search all used different parameter combination, as can be seen from Table 9.3. There was no best configuration across the models, indicating that traits must be considered independently when it comes to training the models

Table 9.3.: Parameters for Baseline Predictor

Personality Trait	**C**	**epsilon**
Extraversion	5	0.2
Agreeableness	20	0.8
Neuroticism	20	0.1
Openness	15	0.2
Conscientiousness	0.1	0.8

Support Vector Regression Results

After establishing the baseline using a linear SVR, the model was trained using LIWC features, with an *RBF* and *sigmoid* kernel. The model was also trained and tested in a similar fashion using TF-IDF unigrams of words, but tested with a *linear*, *RBF* and *sigmoid kernel.* The results can be found in Table 9.4.

Table 9.4.: Score for SVR Run with LIWC and TF-IDF Features

Features	**Extra-version**	**Agree-ableness**	**Neuro-ticism**	**Open-ness**	**Conscien-tiousness**
LIWC	0.06	0.05	0.27	0.29	0.08
TF-IDF	**0.16**	**0.19**	**0.30**	**0.30**	**0.21**

When run using different kernels for the LIWC features, the SVR model performed better for all traits except consciensiousness. However, the results were still poor for extraversion, agreeableness and conscientiousness, with scores indicating a weak positive correlation. Surprisingly, the model was able to achieve a score of 0.27 for Neuroticism with an *RBF* kernel, up from 0.12 using a linear kernel. All the best performing estimators returned from the grid search used *RBF* as the kernel. The results were interesting, showing that the kernel has a substantial effect on the model's ability to capture information. Another interesting finding was that the best estimators all had a *C* parameter of 20. This may suggest that the parameter configuration is more stable when trained on the *RBF* kernel.

Models trained on TF-IDF features performed better than the LIWC features across all traits. While the increase in performance was low for neuroticism and openness, the results show a substantial increase for extraversion, agreeableness and conscientiousness, with an average performance increase of nearly 179.2% across these traits. These results suggest that feature vectors generated from the vocabulary itself are able to pick up more information than the predefined LIWC features.

The SVR model also was tested on sentence embeddings generated from the 2nd, 11th, and 12th layer of ALBERT and the 2nd, 5th and 6th layer of DistilBert.

The model was trained and tested on sentence embeddings from these layers, with and without stopwords in the corpus. The results can be seen from Table 9.5 and Table 9.6.

Sentence embeddings outperformed TF-IDF on every trait except Agreeableness. The increased performance was expected, as transformer-based models are able to capture contextual information, which TF-IDF is not. Sentence embeddings with stopwords included performed on average 21.3% better than the models trained on sentence embeddings with stopwords removed. This suggests that what the stopwords contained in the Gensim vocabulary may not represent the stopwords contained in the *personality*$_d$ corpus. The Openness model was able to achieve 5.3% better performance without stopwords included in the extracted sentence embeddings. In general, the DistilBert generated sentence embeddings yielded the best results, only beaten by ALBERT sentence embeddings for Openness.

Table 9.5.: Score for SVR Run with Sentence Embeddings Containing Stopwords

Features	**Extra-version**	**Agree-ableness**	**Neuro-ticism**	**Open-ness**	**Conscien-tiousness**
ALBERT L2	0.15	0.06	0.34	0.33	**0.25**
ALBERT L11	0.18	0.09	0.32	0.33	0.22
ALBERT L12	0.19	0.12	0.27	0.33	0.21
DistilBert L2	0.17	0.13	**0.36**	**0.36**	0.22
DistilBert L5	**0.22**	0.15	0.24	0.35	0.25
DistilBert L6	0.15	**0.17**	0.31	0.36	0.23

Table 9.6.: Score for SVR Run with Sentence Embeddings Not Containing Stopwords

Features	**Extra-version**	**Agree-ableness**	**Neuro-ticism**	**Open-ness**	**Conscien-tiousness**
ALBERT L2	0.10	0.07	0.29	0.33	0.16
ALBERT L11	**0.17**	0.09	0.29	0.34	0.19
ALBERT L12	0.09	**0.12**	0.25	0.36	**0.19**
DistilBert L2	0.12	0.10	0.30	**0.38**	0.17
DistilBert L5	0.14	0.11	**0.33**	0.37	0.16
DistilBert L6	0.14	0.11	0.31	0.35	0.14

Multilayer Perceptron Results

The Multilayer Perceptron (MLP) was trained and tested on sentence embeddings from ALBERT and DistilBert. The results can be seen in Table 9.7 and Table 9.8. The MLP models performed worse than the SVR models trained on sentence embeddings for all traits except for openness, which obtained a result of 0.39 when run on sentence embeddings without stopwords. The Result came as a surprise as MLPs generally are well suited for high dimensional data instances. Similar to the SVR models trained on sentence embeddings, DistilBert generated features counted for 60% of the best performing models. For MLPs trained on sentence embeddings with stopwords included, there was no single best layer to extract the word embeddings from. For the MLPs trained on sentence embeddings with the stopwords removed, the second-to-last layer of DistilBert proved to be best in general.

Table 9.7.: Score for MLP Run with Sentence Embeddings Containing Stopwords

Features	**Extraversion**	**Agreeableness**	**Neuroticism**	**Openness**	**Conscientiousness**
ALBERT L2	0.12	0.07	0.12	0.22	0.20
ALBERT L11	0.15	0.10	0.25	0.24	**0.20**
ALBERT L12	**0.19**	0.15	0.12	0.19	0.09
DistilBert L2	0.13	**0.16**	0.23	**0.35**	0.13
DistilBert L5	0.17	0.16	0.31	0.35	0.13
DistilBert L6	0.16	0.08	**0.31**	0.35	0.18

Table 9.8.: Score for MLP Run with Sentence Embeddings Not Containing Stopwords

Features	**Extraversion**	**Agreeableness**	**Neuroticism**	**Openness**	**Conscientiousness**
ALBERT L2	0.12	0.08	0.27	0.27	0.15
ALBERT L11	**0.16**	0.02	0.14	0.30	**0.19**
ALBERT L12	0.13	0.01	0.07	0.32	0.10
DistilBert L2	0.11	0.17	0.19	0.32	0.13
DistilBert L5	0.15	**0.18**	**0.28**	**0.39**	0.09
DistilBert L6	0.13	0.12	0.27	0.37	0.07

Gaussian Process Results

The Gaussian Process (GP) models were trained on sentence embeddings generated from ALBERT and DistilBert. The results of the runs can be seen from Table 9.9 and Table 9.10. The results were considered interesting, as the GP models were able to outperform every other model architecture on four out of five traits, only beaten on extraversion by the SVR model using sentence embeddings. As evident by the results from the other models, the sentence embeddings generated from documents with stopwords included proved to perform generally better across all traits. The only exception was openness, which received the best score across all models, with a Pearson Correlation Coefficient of 0.40.

Table 9.9.: Score for GP Run with Sentence Embeddings Containing Stopwords

Features	**Extra-version**	**Agree-ableness**	**Neuro-ticism**	**Open-ness**	**Conscien-tiousness**
ALBERT L2	0.15	0.20	0.35	0.33	0.25
ALBERT L11	0.19	0.21	0.33	0.33	0.21
ALBERT L12	0.19	**0.25**	0.33	0.33	0.21
DistilBert L2	0.17	0.21	**0.37**	**0.36**	0.25
DistilBert L5	0.20	0.20	0.33	0.36	**0.26**
DistilBert L6	**0.21**	0.18	0.32	0.35	0.25

Table 9.10.: Score for GP Run with Sentence Embeddings Not Containing Stopwords

Features	**Extra-version**	**Agree-ableness**	**Neuro-ticism**	**Open-ness**	**Conscien-tiousness**
ALBERT L2	0.11	0.13	**0.33**	0.35	0.21
ALBERT L11	0.15	0.14	0.30	0.34	0.19
ALBERT L12	0.14	**0.16**	0.28	0.34	0.20
DistilBert L2	0.15	0.14	0.32	**0.40**	0.17
DistilBert L5	0.15	0.13	0.33	0.37	**0.21**
DistilBert L6	**0.17**	0.14	0.30	0.35	0.19

Adaptive Boosting Results

The AdaBoost model was trained on sentence embeddings generated from ALBERT and DistilBert. The results from the run can be seen from Table 9.11 and Table 9.12. The results were interesting as the AdaBoost model did not rank particularly high on any of the traits when compared to the other model architectures. This indicated two things; 1) the AdaBoost models were not able to fully learn from past prediction errors, and 2) the AdaBoost models did not curb any over-fitting. The latter suggesting that the risk of over-fitting was low for the $personality_d$ dataset.

Table 9.11.: Score for AdaBoost Run with Sentence Embeddings Containing Stopwords

Features	**Extra-version**	**Agree-ableness**	**Neuro-ticism**	**Open-ness**	**Conscien-tiousness**
ALBERT L2	0.08	0.22	0.32	0.29	0.22
ALBERT L11	0.10	**0.25**	0.33	0.30	0.24
ALBERT L12	0.14	0.23	0.33	0.31	0.19
DistilBert L2	0.11	0.24	**0.34**	**0.34**	0.21
DistilBert L5	**0.18**	0.15	0.29	0.31	0.23
DistilBert L6	0.18	0.19	0.32	0.32	**0.24**

Table 9.12.: Score for AdaBoost Run with Sentence Embeddings Not Containing Stopwords

Features	Extraversion	Agreeableness	Neuroticism	Openness	Conscientiousness
ALBERT L2	0.05	0.12	0.28	0.29	0.21
ALBERT L11	**0.17**	0.07	0.33	0.30	0.22
ALBERT L12	0.16	0.16	0.29	0.30	**0.22**
DistilBert L2	0.14	**0.17**	**0.34**	**0.38**	0.20
DistilBert L5	0.06	0.11	0.26	0.31	0.21
DistilBert L6	0.13	0.09	0.29	0.27	0.16

Top Ranking Regression Models

Figure 9.1 shows the best score achieved from each model, for each trait. The Gaussian Process models ranked highest across all traits except extraversion, having an average trait score of 0.30 across all trait. AdaBoost and SVR came in on a close second with an average score of 0.28, followed by MLP with an average score of 0.25. One interesting finding was that the SVR, AdaBoost and MLP architectures all had one trait in which the models scored the lowest. GP was the only architecture with a steady high performance across all traits when compared to the other architectures.

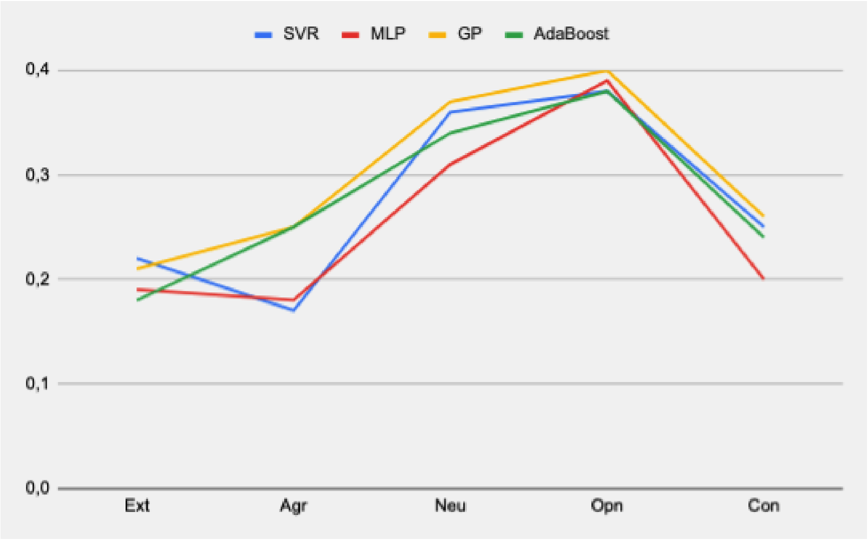


Figure 9.1.: Trait Scores for each Model

Table 9.13 gives a summary of the best performing models, along with their features. As can be seen from the result, sentence embeddings generated by DistilBert were able to capture more contextual information than those generated by ALBERT. The difference was marginal, with an average performance 4.76% higher for DistilBert over ALBERT, when comparing the best models.

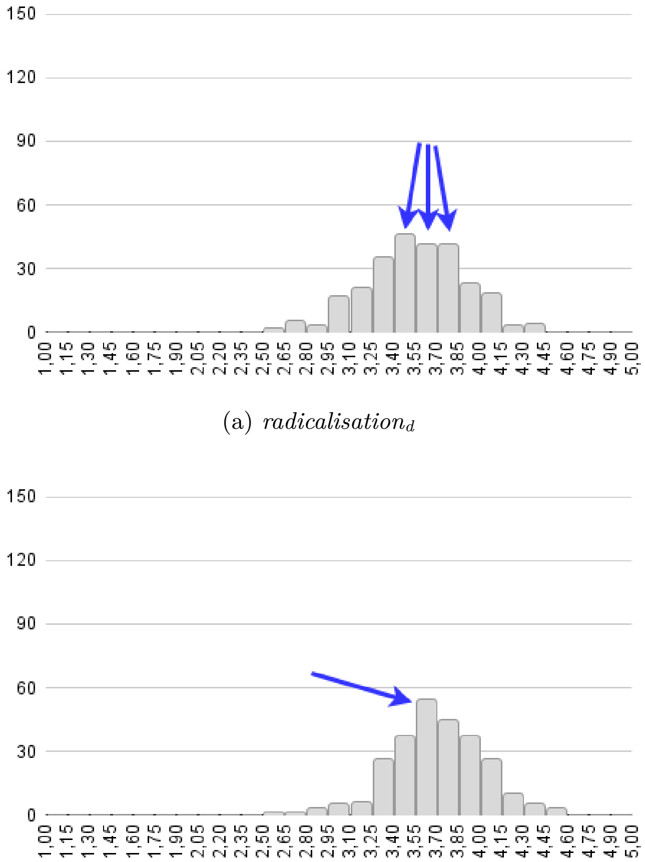
No single layer stood out as giving more contextual information than the others, with an even split between layer 2 and layer 5 for DistilBert. Another interesting finding was that a total of seven openness models, using sentence embeddings without stopwords, ranked higher than the best performing model using sentence embeddings with stopwords. This was surprising, as sentence embeddings with stopwords performed on average 22% better for the best performing models, for the remaining traits. A complete summary of the hyperparameters used in the best performing feature-model combination can be found in Appendix H.

Table 9.13.: Summary of Best Performing Models with Features

Trait	**Score**	**Model**	**Feature**	**Layer**	**Stopwords Removed**
Extraversion	**0.22**	SVR	DistilBert	L5	
Agreeableness	**0.25**	GP	ALBERT	L12	
Neuroticism	**0.37**	GP	DistilBert	L2	
Openness	**0.40**	GP	DistilBert	L2	**x**
Conscientiousness	**0.26**	GP	DistilBert	L5	

9.3.2. Experiment 2 - Results

For Experiment 2, the best performing models from Experiment 1 was selected. In order to keep a uniform format for the pre-processed datasets, stopwords were kept for all features, despite the Openness-model scoring higher with stopwords removed. Keeping the stopwords for all features reduced the average performance across the traits from 0.30 to 0.292, which was considered acceptable by the author. Figure 9.2 to 9.6 gives a side-by-side comparison of the predicted trait scores, with the results from $\textit{radicalisation}_d$ on the left and $\textit{non-radical}_d$ on the right. A summary of the statistics for the predicted traits is given in Table 9.14 and Table 9.15.

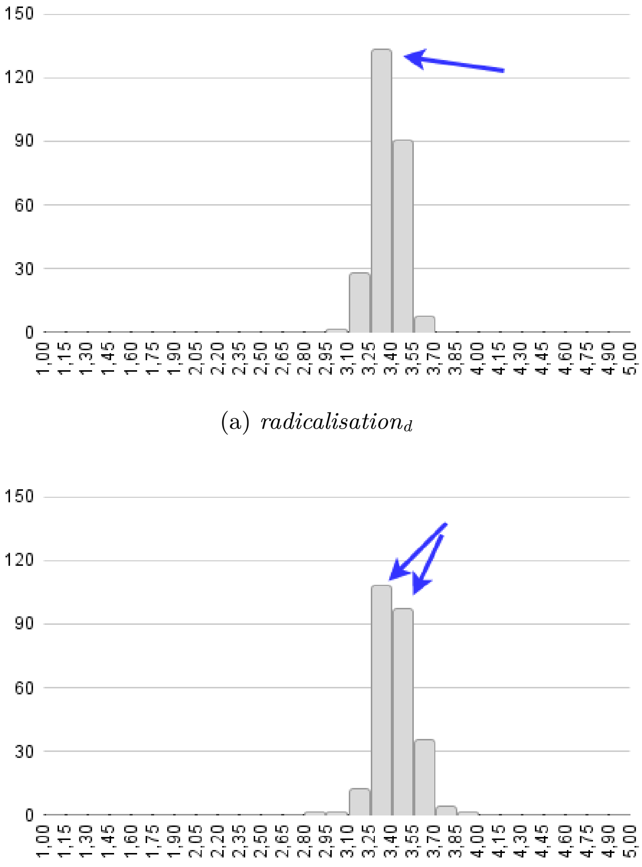


(a) *radicalisation$_d$*

(b) *non-radical$_d$*

Figure 9.2.: Predicted Scores for Extraversion

The results for the extraversion prediction yielded the highest difference in predicted values for the two datasets, with the average predicted score of *non-radical$_d$* being 0.16 higher than for *radicalisation$_d$*. Both predictions had a normal distribution, as can be seen from Figure 9.2. While *non-radical$_d$* had a peak around 3.59-3.75, the majority of the predictions for *radicalisation$_d$* were evenly distributed around 3.25-3.75. This was considered interesting as the extraversion prediction achieved the highest standard deviation for both datasets.

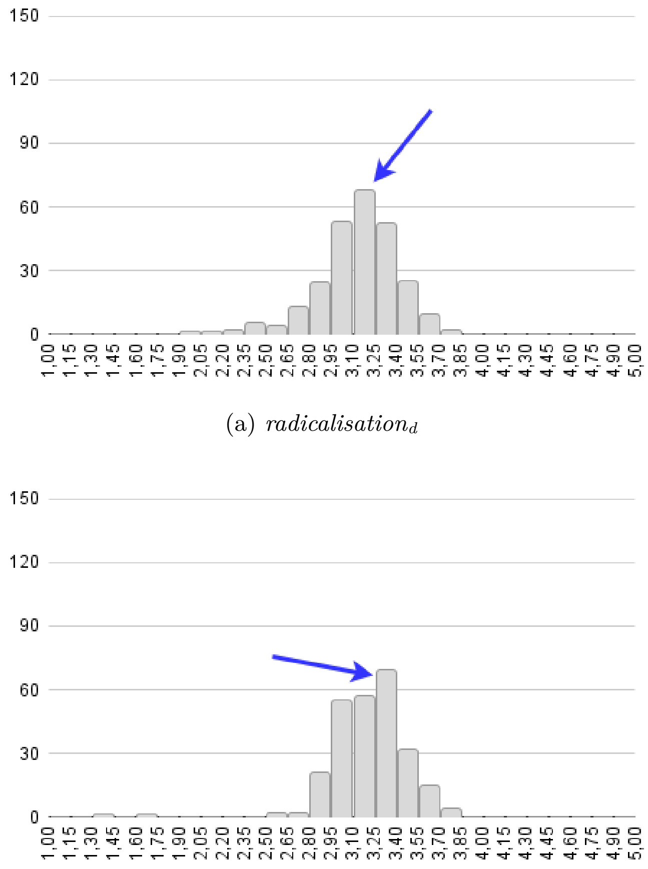


(a) *radicalisation$_d$*

(b) *non-radical$_d$*

Figure 9.3.: Predicted Scores for Agreeableness

The results for the agreeableness prediction can be seen from Figure 9.3. The predicted scores for *radicalisation$_d$* had a clear peak around 3.25-3.40. The predicted scores for *non-radical$_d$* was evenly distributed around 3.25-3.55. Despite the difference in distribution, the average predicted score only differed by 0.05 across the datasets, while also yielding the lowest standard deviation across all traits.

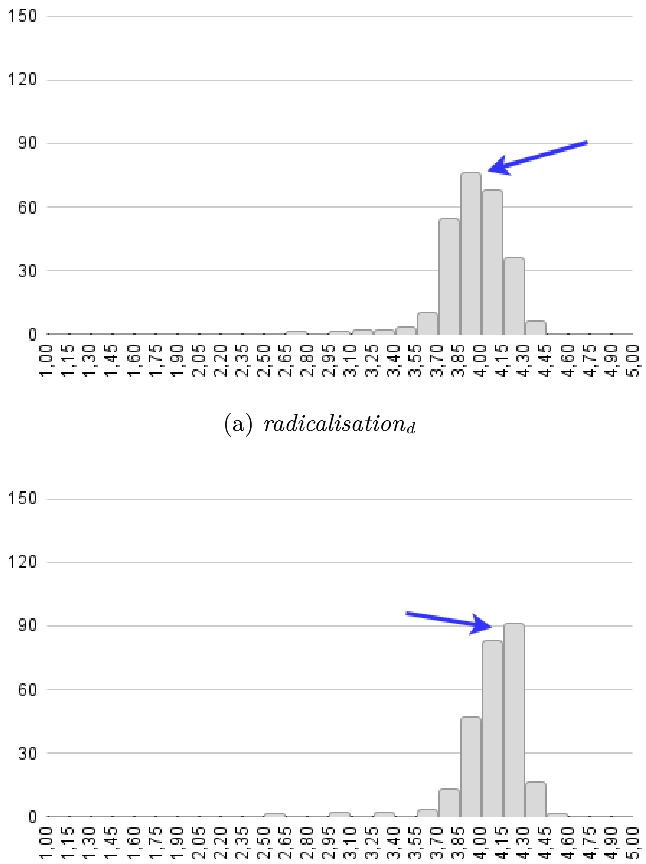


(a) *radicalisation*$_d$

(b) *non-radical*$_d$

Figure 9.4.: Predicted Scores for Neuroticism

The results for the neuroticism prediction can be seen from Figure 9.4. The predicted scores for both datasets had a fairly normal distribution around the same values, with both predictions having multiple outliers to the left of the curve. There was close to no difference in the predictions between the two datasets. Similar to that of agreeableness, the datasets only differed by 0.05 for the average predicted score, and 0.02 for the standard deviation.

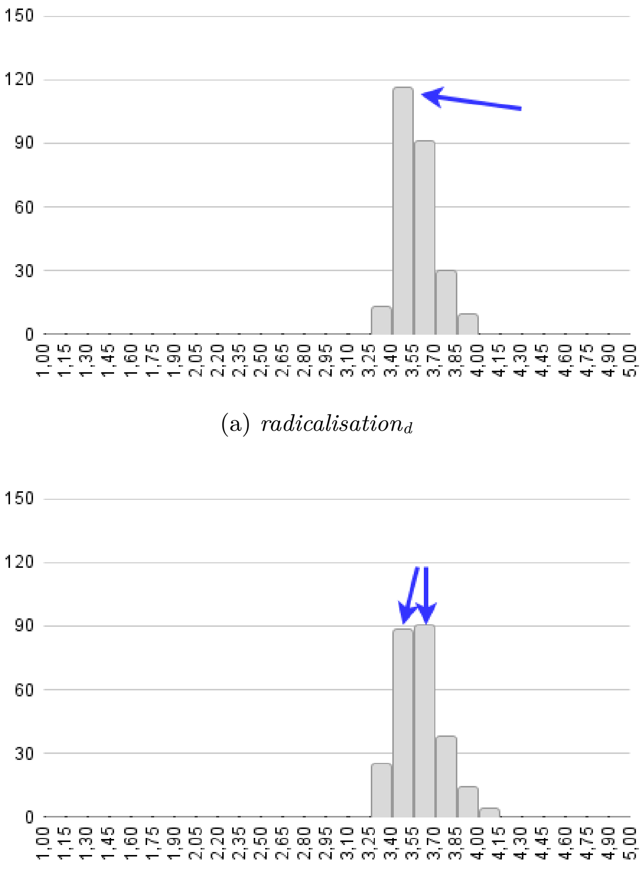


(a) *radicalisation$_d$*

(b) *non-radical$_d$*

Figure 9.5.: Predicted Scores for Openness

The results for the openness prediction can be seen from Figure 9.5. The predicted scores for *radicalisation$_d$* had a normal distribution around the values 3.40-4.45, with some outliers to the left. The predictions for *non-radical$_d$* was skewed to the left. Similar to that of *radicalisation$_d$* the distribution had several outliers to the left. The openness prediction achieved the second highest difference in average prediction score, with only 0.13 separating *radicalisation$_d$* and *non-radical$_d$*. Despite the difference in average values, the distributions bear a close resemblance to each other while being the only distribution with equal standard deviation.



(a) *radicalisation*$_d$

(b) *non-radical*$_d$

Figure 9.6.: Predicted Scores for Conscientiousness

The results for the conscientiousness prediction can be seen from Figure 9.6. There was close no to difference between the two predictions, with only 0.01 separating the average predicted values of *radicalisation*$_d$ and *non-radical*$_d$.

Table 9.14.: Statistics for *radicalisation*$_d$ Prediction

Trait	**Max**	**Min**	**Avg**	**Std**
Extraversion	4.44	2.54	3.54	0.34
Agreeableness	3.66	3.03	3.37	0.10
Neuroticism	3.82	2.01	3.14	0.28
Openness	4.43	2.75	3.95	0.22
Conscientiousness	4.00	3.28	3.58	0.13

Table 9.15.: Statistics for *non-radical*$_d$ Prediction

Trait	**Max**	**Min**	**Avg**	**Std**
Extraversion	4.58	2.58	3.70	0.32
Agreeableness	3.87	2.86	3.42	0.13
Neuroticism	3.81	1.31	3.21	0.26
Openness	4.49	2.52	4.08	0.22
Conscientiousness	4.06	3.31	3.59	0.15

While the predictions for *non-radical*$_d$ yielded a higher average predicted score across all traits when compared to *radicalisation*$_d$, the difference between the two datasets was considered marginal.

9.3.3. Experiment 3 - Results

The results from Experiment 3 can be seen from Table 9.16. Neuroticism and Frustration yielded the highest score, with a positive correlation of 0.27. This was not surprising as neuroticism represents the emotional stability of an individual. One interesting finding was that neuroticism ranked high in terms of correlation across all indicators, when compared to the other traits. This was considered interesting, suggesting a prevalence of neurotic traits among people vulnerable to radicalisation. Another interesting finding was that the use of terminology promoting and glorifying jihadism did not seem to correlate notably with any of the indicators. The highest correlation was found for Neuroticism, with a weak negative correlation of 0.12.

Conscientiousness did not seem to correlate notably with any of the indicators. The highest correlation was found for the *negative words* indicator, yielding a weak negative correlation of 0.13.

Openness correlated positively with the use of negative words and negativity towards western society. This was considered interesting, as it conformed with the findings of Wiktorowics (2005), suggesting an openness towards new worldviews among individuals being radicalised.

Table 9.16.: Correlation Between Indicators and Traits

Indicator	**Ext**	**Agr**	**Neu**	**Opn**	**Con**
Indicator 1 - Frustration	0.09	0.00	**0.27**	0.00	-0.05
Indicator 2 - Negative Words	0.13	-0.16	0.16	**0.21**	-0.13
Indicator 3 - Discrimination	0.16	0.06	0.16	-0.03	-0.03
Indicator 4 - Negative West	0.06	-0.15	**0.21**	**0.23**	-0.07
Indicator 5 - Pro Jihad	-0.13	-0.11	-0.12	-0.06	-0.06

10. Evaluation and Discussion

This chapter presents an evaluation and discussion of the research done as part of this Thesis. The chapter is divided into two main sections. In the first section a discussion of the choices leading up to and made during the experiments of this Thesis is presented. This also includes a discussion of the results in light of previous work and state-of-the-art within the field of personality prediction and radicalisation research. The section concludes by discussing limitations and ethical concerns regarding this Thesis. The second section presents an evaluation of the findings of this Thesis with regards to the research questions and research goal are presented in Section 1.2.

10.1. Discussion

Chapter 9 presented the plan, implementation and results of three experiments aimed at answering Research Question 3 and 4, presented in Section 1.2. A total of five datasets were used for conducting these experiments, all of which are presented in Chapter 7. Three existing datasets were selected for training and testing the performance of a selection of personality classification models. Two datasets were manually collected from Twitter, one consisting of users believed to be vulnerable to radicalisation and one of what was considered to be ordinary non-radical users. The two manually collected datasets and the findings from the experiments are considered the main contributions of this Thesis. This section provides a discussion regarding the choices made for answering the reasearch questions, limitations concerning these choices and ethical concerns regarding personality prediction. First, a discussion of the process concerning the manual collection and annotation Twitter data is presented, aimed specifically at Research Question 1 and 2. The second section discusses the choice of personality datasets for the task of building a personality prediction model. The third section discusses the pre-processing steps taken for all the five datasets, and how this might impact the performance of the personality prediction models, along with a discussion of the implementation and performance of the resulting models. The forth section discusses the results of Experiment 2 in light of previous studies on radicalisation and how these results may give insight into the traits of people believed to be at risk of radicalisation. The final section covers ethical concerns regarding the process and experiments conducted as part of this Thesis.

10.1.1. Choice of Personality Datasets

The three datasets used for training the personality prediction model in this Thesis are all described in Chapter 5 and further elaborated in Chapter 7. All datasets had been used in previous task for personality prediction or classification and yielded good results. Two of these datasets, namely myPersonality Twitter Dataset and PAN 2015 Author Profiling dataset, were generated from Twitter data. Stream-of-Conscientiousness was based on essays written by students, but due to the informal nature of the texts written, it was considered to be of a format consistent with the Twitter-based datasets.

One of the challenges encountered when working with the Twitter datasets was the distribution of scores for the personality traits. As can be seen from Figure 7.1 and Figure 7.2, the distribution was normalised for most of the traits, giving few instances with low and high scores for the regression models to train on. The personality prediction being a regression task made the options of over- and undersampling difficult. Also, due to the sparsity of the datasets, undersampling would result in too few instances for the models to train on, while oversampling could increase the risk of overfitting the data. Keeping the distribution as it is was thus considered a lesser of two evils but could still have impacted the performance of the models.

For stream-of-conscientiousness, another challenge presented itself as all personality traits were represented as binary values, meaning the trait was either present or not. The dataset was still considered to have useful information for the models to train on and it was desirable to keep the dataset. As this Thesis aimed to train models for prediction, the values would have to be converted to numerical values. The approach chosen is presented in Section 7.4.3. The approach yielded promising results in the workings of Nordnes and Gran (2019) who worked with the same dataset in their personality prediction study. One aspect worth considering with this approach is the risk of overfitting in the models. Each binary value was represented with a constant numerical value, which may worsen the generalisation capabilities of the trained models. One option would be to keep the binary values for each dataset and train the models for classification. This could however affect the level of insight into the personality traits of the individuals found in $radicalisation_d$ and $non\text{-}radical_d$ and may therefore not have been suitable for the research goal of this Thesis. However, experimenting with other datasets or classification tasks could be examined in future research.

Another challenge that presented itself was the sparsity of data. Due to privacy concerns, the myPersonality Facebook Dataset was made unavailable by the provider as of 2018. The dataset has for long contributed to state-of-the-art results due to its size and level of information. The datasets used in this Thesis were carefully selected to fit the domain and contain enough information to gain reasonable performance from the regression models. However, it is worth noting that the sparsity of available data could have affected the performance of the final prediction models. There is a need for a new gold standard dataset for personality prediction and classification, and future work should therefore explore the possibilities of obtaining such a dataset.

10.1.2. Collection and Annotation of Data

In order to reach the goal of this Master's Thesis, two datasets had to be manually collected, one consisting of users believed to be vulnerable to radicalisation and one representing your everyday Twitter user. Section 7.2 briefly describes the challenges when it came to building the $radicalisation_d$ dataset. The first challenge was to identify individuals on Twitter believed to be vulnerable to radicalisation. There were no existing datasets known to the author of this Thesis, and such a dataset would therefore have to be manually or semi-manually constructed. Due to a lack of studies on Twitter behaviour prior to radicalisation, a data collection methodology would have to be derived. The proposed methodology was based on literature studied as part of this Thesis and related work of radicalisation on Twitter. The initial idea of this Thesis was to harvest existing datasets from previous studies on radicalisation and use these datasets to target individuals at risk of being radicalised. A total of four datasets were believed to be of interest for this study and requested from the authors. None of the authors were able to provide the datasets, due to either having left their positions in academia or internal restrictions forbidding them to share the data. However, upon exploring the *How ISIS Uses Twitter* dataset, it was found that roughly 95% of the users were suspended. This illustrated one of the challenges when working with radicalisation on Twitter. It was found that the requested datasets would most likely have a high percentage of suspended accounts and thus not be useful.

The first step of building a dataset was to identify users believed to be radicalised. Due to a high suspension rate, the users would have to be identified over a short time span. The users were identified through two Twitter initiatives, explained in Section 7.2.2. The approach was suggested by the author of this Thesis and founded in the guidelines proposed by Parekh et al. (2018), and resulted in a dataset of 121 users. As of 22.01.2022, almost three months after the dataset was created, a total of 55 users, or 45.5%, had been suspended. For comparison, Conway et al. (2019) found that 65% of pro-ISIS users and 20% of jihadist users were shut down by Twitter within 70 days. This Thesis did not aim at identifying users affiliated specifically with ISIS, but instead Islam-related jihadism. The high suspension rate of the 121 users identified may therefore suggest that the correct users were identified.

Another challenge when identifying radical users was the content which the annotators were exposed. At the time of writing this Thesis, Twitter do not provide any service to manually verify content before it is published. Instead, content in violation of their policies is continuously reported and removed. This allows for disturbing content to be open to the public until removed. Exposure to disturbing content is a serious issue that is hard to get around when working with radicals on social media. Both annotators of the dataset found it emotionally draining working with tweets due to the graphical and textual content contained in the tweets. The annotators did not account for the strain of having to go through this material, which caused the data collection and annotation phase to take longer than expected. Even though the annotators did not sustain any long-term effects from the annotation process, it is worth considering the effects of working with radical content for future research projects.

The lack of empirical studies on radicalisation paths posed a challenge for deriving a annotation scheme. The resulting scheme and choice of indicators was based on overlapping consensus between literature. Relevant users were first sampled by using a dictionary representing the indicators, before being manually verified, categorised and annotated. The results from the process was summarised in Table 7.2. Around 28% of the sampled followers had terms matching the constructed dictionary. This was a high percentage, but still filtered out around 72% of the collected followers. After manually verifying and annotating a random sample of 10% of the users matching the terms, only 30.6% were labelled as vulnerable towards radicalisation. This illustrates the findings from reviewing the work of Rowe and Saif (2016) and Fernandez et al. (2018). Relying on a purely lexical approach for identifying radicals or users vulnerable to radicalisation may not be sufficient, as you run the risk of including irrelevant users. Another challenge was the separation between endorsement and dissemination of information. For a user to be labelled as glorifying or supporting jihadist ideals, there would have to be a clear incentive behind the tweet, retweet or reply. This included a risk of falsely annotating users as vulnerable to radicalisation who in fact may already be radicalised. Previous studies had however pointed out the distinction between endorsement and dissemination, further exemplified by the faulty take-down of Twitter accounts by Anonymous, explained in Section 7.2. The cost-benefit of this distinction was reviewed by the author, was considered to be an important point.

Unrelated was by far the most represented category during the annotation process. Considering the sampled 10% as a representation of the 9,986 users matching the dictionary, users believed to be vulnerable to radicalisation only made up 4.2% of the collected followers. In comparison, Rowe and Saif (2016) used a similar approach and found around 1% of their users to be radical. Based on the theory presented in Chapter 2, not everyone at risk of radicalisation end up radicalised. It was therefore expected that the number from the final annotation would to be higher.

One aspect to consider with regards to the annotation and defined indicators, is to which degree they capture the radicalisation process. The point of the indicators defined in Section 7.2.3 was to capture behaviour that may not unfold as radicalisation, but indicates whether an individual is susceptible towards radical ideas. Not every indicator of radicalisation is observable online. For instance, is an indicator of ongoing radicalisation or susceptibility towards radicalisation that an individual inquires about jobs that provide sensitive access. Non-observable indicators may therefore increase the risk that certain users are neglected during annotation. Also, radicalisation is not a linear path, so the proposed methodology is not guaranteed to find everyone at the brink of radicalisation. Only 48.3% of the collected followers had 40 or more tweets and was matched against the lexicon. There might be lone-wolfs and non-vocal people who were not captured by the lexical and manual verification. Debating or publishing content affiliated with radicalisation is not a guaranteed attribute for those undergoing it. There is therefore a highly probable that the Thesis left out several users which in fact may be at risk of radicalisation.

Another aspect to consider is to which degree the lexical approach gives a good representation of terms affiliated with radicalisation. The terms chosen for this Thesis and defined in the composed dictionary were based on previous studies on radicalisation. The most common terms used among the users found in $radicalisation_d$ was depicted in Figure 7.4. From the word cloud it was apparent that few terms from the actual dictionary were present. Still, the terms contained in the word cloud appeared to be oriented around a few common themes; radical movements, religion, foreign western countries, war, government and Middle-Eastern countries. Several of these themes were covered in the dictionary by other terms, which may suggest that the indicators are correct, but the dictionary should be expanded. Despite there being similarities in terms of the indicators themselves, the comparison reiterates the importance of manually verifying data from lexical approaches, as the terms themselves differed somewhat.

In order to predict personality traits of individuals at risk of radicalisation, all retweets were left out. Retweets were however kept in the initial annotation process. This posed a challenge as the Twitter API does not provide an efficient solution to extracting tweet count for users, with retweets discarded. The result was that several of the users having 40 or more English tweets were in fact retweets that could not be used in the final prediction. Of the 305 identified users, 27 users had to be removed due to the lack of content written in the users own words. In order to overcome this issue, one would have to have a script manually check the tweets of each user, verify the language and count the frequency of tweets, replies, quotes and retweets. Considering the limits on the number of request to the API, imposed by Twitter, this would not be possible considering the time frame of this Thesis.

10.1.3. Pre-Processing of Data

The pre-processing steps applied in this Thesis were described in Section 7.4-7.6. The data was processed to obtain a uniform format suitable for feature extraction. Despite being considered necessary, these modifications could cause semantic or contextual information to be lost from the original datasets. The conversion of mentions and URLs to the abbreviated *MENTION* and *URL* could potentially cause a loss of information. URLs may point to specific resources that provide additional information on the topic of a tweet. Capturing the content of such URLs may be complex and was outside the scope of this Thesis but could potentially give extra information in terms of generating a good feature representation.

One aspect worth considering was the concatenation of tweets per user. The concatenation was a result of the datasets being annotated on a user-level. While it is not uncommon to concatenate documents once annotated on a user-level, there are aspects worth considering when doing so. Tweets may be contextually independent of each other, meaning that a certain tweet may not be related to another tweet from the same user. Concatenating these tweets and extract contextual features may result in proportions of the context being lost. Another aspect with regards to concatenation is the limitations imposed by the transformer-based architectures. Both ALBERT and DistilBert take in a maximum of 512 tokens when generating word embeddings. Concatenated documents whose tokenisation were longer than 512 tokens would then be truncated, resulting in a loss of information. One way to cope with this would be to split tokenised documents with a size greater than 512. This option was not considered as the average number of words per concatenated document was close to the limit imposed by the architectures. However, due to the sparsity of data for training the personality prediction models, future research could consider either splitting the tokenised documents or avoid concatenating the tweets.

Another aspect worth considering was the selection of abbreviated words. While abbreviations are not kept as a standard part of ALBERT and DistilBert's vocabulary, only a handful of abbreviations were selected for this Thesis. The selection was based on two factors;

- It should be a common abbreviation used on social media
- here should be a low risk of falsely interpreting a term as an abbreviation

For instance, could the terms *r*, *y*, and *n* be abbreviated forms of *are*, *yes*, and *no*, but they could also be the letters themselves. This would depend on the context, and several known abbreviations were therefore discarded. It is however worth noting that for some instances, neglecting certain abbreviations could result in loss of information with regards to the transformer-based feature extraction.

10.1.4. Building Personality Prediction Model

A total of three datasets were used for training a model for personality prediction. Several models were available to the author, where a handful were selected based on previous research reviewed as part of Chapter 6. In order to derive the best performing models, a search for finding optimal parameters and features was conducted as part of Experiment 1. The Gaussian Process architecture performed best for the majority of traits, as can be seen from Table 9.13. Despite being the best architecture, none of trait models shared the same parameter configuration. Extraversion was the only trait with SVR as the best performing architecture. Due to the traits needing separate models for prediction it was decided to select the best performing models for Experiment 2, except for the Openness model. A total of seven Openness models with stopwords removed outperformed the best performing model with stopwords included. In order to keep a uniform pre-processing of the data, it was decided to select best performing Openness model with stopwords included. This only reduced the average test score from 0.300 to

0.292 but may however have impacted the final Openness prediction for *radicalisation*$_d$ and *non-radical*$_d$.

While the majority of the selected models used DistilBert-generated features, no single layer stood out as giving more contextual information across the traits. Similar to that of selecting the models, the single best feature representation was selected for each trait. One option would be to choose the features, models and parameters with the best average performance to maintain a uniform prediction for Experiment 2. Due to the variance in performance across the implementations and needing one separate model per trait, it was decided to choose the best performing feature-model configuration for each trait.

There were several aspects to consider with regards to the performance of the models selected for Experiment 2. Regression tasks yields on average a lower performance than that of classification tasks. As was evident from the reviewed work in Chapter 6, personality prediction is no exact science. What is considered an acceptable performance of the models thus depend on the prediction task itself. The models selected for Experiment 2 achieved a reliable average accuracy of 0.292 across the traits. The best performing features and parameters selected for Experiment 2 may not have been the optimal combination for the task but produced the best result across all the tested configurations. In reviewing related work, several tasks using word embeddings as features came out on top. For comparison, LIWC and TF-IDF were tested in combination with SVR but were beaten on most traits. There may have been other parameter and features that could have produces better results, but due to time restrictions and personality prediction being only a part of this Thesis, it was decided to go with these models.

Another aspect to consider when reviewing the performance of the models was the amount of available data to train on. Of the reviewed work, presented in Chapter 6, myPersonality Facebook was the most popular dataset across the studies. It was for long considered a gold standard for personality prediction and classification, much due to the rich amount of data to train on. When comparing the sizes of the datasets used for this Thesis and the amount of data used in previous studies utilising the original myPersonality dataset, it was expected that the models would achieve considerably lower scores. Implementing Experiment 2 as a classification problem rather than a regression problem may have improved upon the performance of the models. However, this may have resulted in less insight into the score distribution of traits for *radicalisation*$_d$ and *non-radical*$_d$. Converting to classification was therefore not considered.

When comparing the models trained in this Thesis with related work using regression models and sparse amounts of data, the models places themselves somewhere in between. Preotiuc-Pietro et al. (2016) achieved an average Pearson Correlation Coefficient of 0.25 when predicting Dark Triad personality traits of Twitter users. The accuracy was seen as reliable and the results good, considering the regression models were trained on only 491 users.

When comparing with state-of-the-art performances, the most relatable work was produced by Arnoux et al. (2017) who at the time were able to outperform state-of-the-art results with eight times fewer input data. Similar to that of this Thesis, their approach used word embeddings as features, in combination with Gaussian Process Regression, to predict Big5 personality traits of Twitter users. Their best performing models scored an average Pearson Correlation Coefficient of 0.33 across all traits. Similar to that of this Thesis, their performance of the Extraversion model ranked lowest, followed by Agreeableness and Conscientiousness. One interesting finding was that several of the Openness models of this Thesis were able to outperform those of Arnoux et al. (2017).

Another aspect worth discussing was the final predictions conducted as part of Experiment 2. At first glance the results of the experiment may suggest few differences in the predicted traits of *radicalisation*$_d$ and *non-radical*$_d$. Despite the distribution being within the same value are, the distribution itself differs somewhat between the two datasets. For instance, did Extraversion have a clear peak around 3.55-3.70 for *non-radical*$_d$. For *radicalisation*$_d$ there did not appear to be a clear a peak. Instead the distribution was evenly distributed around 3.40-3.85, with a small peak around 3.40. This theme was evident for the majority of traits, suggesting that there may in fact, though marginal, be differences between the two datasets in terms of personality traits. These differences may however also be the result of the models themselves. A better performing model may produce different results, showing either differences or similarities between the two datasets. Another factor that may contribute to the results is the size of the datasets used in this Thesis. As already discussed, the performance of the models may have been increased with a larger dataset. Also, increasing the size of *radicalisation*$_d$ and *non-radical*$_d$ may contribute more information about any differences between the two domains. As was evident from Experiment 3, there exists interesting correlations between the traits themselves and the indicators defined in Section 7.2.3. For instance were there a clear correlation between Neuroticism and the *Negative West* and *Frustration* indicator. Similarly were there a clear correlation between Openness and the *Negative West* and *Negative Words* indicator. Future research should aim to investigate the effects of increasing the size of the datasets, both for the prediction task and for training the models, to unveil any larger differences between the two domains explored in this Thesis.

10.1.5. Ethics

The data used as part of this Thesis consist of opinions, emotions, thoughts and information expressed by real individuals online. While the datasets used for training and testing the performance of the personality prediction models were provided willingly by the individuals, the *radicalisation*$_d$ and *non-radical*$_d$ were collected without knowledge or permission from the individuals contained in the dataset. This raised ethical concerns that were important for the author to consider. In order to ensure that these individuals were not exposed, the data was only handled by the author and a second annotator used for measuring the inter-annotator agreement. Upon passing the data to the second annotator, the users were anonymised by converting the usernames to randomly generated IDs. It is however worth noting Twitter's built-in search functionality provides the opportunity to search for exact words or phrases, allowing users to identify users based on their tweets. In order to ensure that individuals could not be identified, all tweet examples presented in this Thesis were fictional, but based on common language expressed in the datasets. As part of ensuring the privacy of the users examined, a binding contract was installed with the second annotator, forbidding the sharing of information from the datasets.

It is however worth noting the cost-benefit of these ethical concerns in light of privacy versus security. The workings of this Thesis are meant to provide insight into the personality traits of individuals at risk of radicalisation. This insight could in turn prove useful for identifying individuals prior to them being drawn into radical communities. Whether these individuals share certain personality traits or not may therefore give insight into how they can and cannot be identified.

Another ethical aspect, discussed in Section 10.1.2, was the strain of annotating radical content. This is a serious issue that is hard to get around when working with radical communities on social media. Future work should therefore aim to incorporate plans to account for any possible after-effects of being exposed to disturbing content.

10.2. Evaluation

The goal of this Master's Thesis was to investigate the personality traits of Twitter users believed to be at risk of radicalisation. In order to best reach this goal, four different research questions were formalised. These research questions were meant to serve as guide and help structure the research to derive the main goal. The following section will evaluate the findings of this Thesis with regards to the research questions and the main goal.

Research Question 1 *What are indicators of individuals being radicalised on social media?*

In order to reach the goal of this Master's Thesis, an understanding of the domain had to be established. As discussed as part of Section 1.1, much research has been devoted to understanding radicalisation and radicals both in society and on social media. As part of the literature review, several studies on both radicals, models for radicalisation, and prediction and classification tasks using machine learning were explored and presented as part of Chapter 2 and Chapter 6.

While several previous prediction and classification studies had focused on individuals already radicalised, this research question focused on building an understanding of the contributing factors leading up to becoming radicalised. Several models and indicators were explored and presented in Section 2.2. As was evident from the review, much of the focus within radicalisation research were on the process of radicalisation rather than the individual itself. Radicalisation is not a linear path followed by every individual, but instead a series of steps and turns which may unfold differently for everyone. Many of the reviewed models were conceptual, rather than empirical, some conflicting and some based on common consensus between researchers. In order to derive a set of indicators, the focus was put on consensus between researchers. A total of five indicators were selected and presented as part of Section 7.2.3. While previous studies such as Fernandez et al. (2018) and Rowe and Saif (2016) had focused on one single indicator, namely positive ideas towards jihad, this Thesis proposed four new indicators for identifying individuals at risk of radicalisation. The four indicators were based on common steps within the radicalisation models presented in Section 2.2 and a study by Lara-Cabrera et al. (2017) measuring the correlation between different indicators and users found in the *How ISIS Uses Twitter* dataset. The indicators were used both in the lexical approach presented in Section 7.2.1 and as part of the final annotation scheme presented in Section 7.2.3.

The occurrence of each indicator in the tweets of $personality_d$ are shown in Figure 10.1. The occurrence of the indicators differs, with *Frustration* being ten times more prevalent than *Discrimination*. The most frequent indicator among the users identified was the unrelated category, which was expected, as the defined categories may not cover every aspect of the radicalisation process, as discussed in Section 10.1.2.

Table 10.1.: Occurrence of Indicators in Tweets for $personality_d$

Frustration	**Negative Words**	**Discrimination**	**Negative West**	**Pro-Jihad**	**Unrelated**
1020	222	88	146	426	2815

Defining what are good indicators of vulnerability towards being radicalised cannot be based purely on the occurrence of each indicator. As was evident from Table 9.16, both *Negative Words* and *Negative West* had the strongest correlation with the Openness personality traits. This result coincides with the findings of Wiktorowics (2005) who found openness to new world views to be a prevalent factor among individuals being radicalised. Surprisingly, Frustration being the most prevalent of the five indicators yielded almost no correlation with any of the personality traits except for Neuroticism, which yielded the highest overall correlation. Pro-Jihad had no strong correlation with any of the predicted traits. This was also surprising, considering the use of pro-jihad terms was the most used indicator in several of the reviewed classification and prediction studies. While the indicator was considered important, the results show that other indicators may be better suited or provide additional information for identifying individuals at risk of radicalisation. The *Discrimination* indicator did not correlate notably with any of the personality traits, but still yielded the highest correlation for

Extraversion. Considering *Discrimination* was the least prevalent indicator among all the users in $personality_d$ and a low overall correlation, the results may suggest that this was one of the worst performing indicators. However, the low occurrence of *Discrimination* may also have inflicted the correlation calculation by there being to sparse data to make accurate predictions. When comparing this to the literature reviewed, the results conflict somewhat for *Discrimination.* Lara-Cabrera et al. (2017) found Discrimination to had one of the strongest similarities with the fewest outliers among the radical users analysed as part of their study. However, their study used looked at the *How ISIS Uses Twitter* dataset, consisting of already radicalised users. This may suggest that the *Discrimination* indicator is better suited as an indicator of individuals already radicalised.

Not every indicator of radicalisation is observable online. With regards to Research Question 1 a total of five indicators were selected based on the reviewed literature. Based on the finding of this study, *Frustration*, *Negative Words* and *Negative West* show interesting correlations with the predicted personality traits, proving their ability as indicators for vulnerability towards radicalisation. *Pro-Jihad*, while being important for capturing a user's interest in jihadist ideals, proved low correlation with the predicted traits of the users. *Discrimination* having low correlation provided little insight into the traits of the users. The low occurrence may also suggest that the indicator may not be the best for identifying individuals at risk of radicalisation.

Research Question 2 *How can indicators of radicalisation be modelled for building datasets of users vulnerable to radicalisation on social media?*

Research Question 2 is a continuation of Research Question 1 with the aim of manually building a dataset of users vulnerable to radicalisation, which is considered one of the main contributions of this Thesis. The quality of the resulting dataset is dependent on reliability, consistency and validity in both the collection and annotation phase. In order to achieve this, the indicators had to be modelled to derive a set of criteria for collection and annotation.

In order to effectively identify relevant users and reduce the number of irrelevant users for annotation, the indicators found as part of Research Question 1 were used both in the collection and annotation phase. The data collection phase was influenced by the reviewed literature, particularly the works of Fernandez et al. (2018), Rowe and Saif (2016) and Parekh et al. (2018). Parekh et al. (2018) reviewed previous attempts on identifying radical communities on social media to derive a set of guidelines, serving as an inspiration for this Thesis. The data collection consisted of several steps, each set to filter out irrelevant users before the final annotation. The process was thoroughly described in Section 7.2, and consisted of the following steps: 1) Identify radical Twitter users to be used as seed accounts, 2) Collect followers of identified users, 3) Comprise a dictionary of terms related to radicalisation and match with tweets from followers, and 4) Manually annotate a sample from users matching dictionary.

The last step of the data collection process, namely manually annotating the collected users, was considered necessary. This importance of manual verification was made clear as only 305 of the 998 users matching the comprised lexicon turned out to be annotated as vulnerable to radicalisation. The final dataset, referred to as *personality*$_d$ made up 4.2% of the collected followers. When comparing this to the results of Rowe and Saif (2016), as discussed in Section 10.1.2, who found roughly 1% of their users to be radical, the results of the proposed scheme seems reasonable. As discussed in Chapter 2, not everyone who embarks on the radicalisation path turn out radicalised. The number of identified users was thus expected to be higher than those already radicalised.

One minor challenge that presented itself, as discussed in Section 10.1.2, was the separation between dissemination of information and endorsement. For the users to be labelled according to the *Pro-Jihad* category, there had to be a clear incentive behind the tweet. This invited subjective interpretation into the annotation. Despite the challenge related to interpretation, the inter-annotator agreement, presented in Section 7.2.3, yielded a kappa of 0.76 or *substantial* for the tweet categorisation. As discussed in Section 4.2.3, one of the drawbacks of Cohen's Kappa is its tendency to underestimate inter-agreement in cases were there are minority classes. The annotators agreed upon the *not present* or 0 label 2,444 times for the sampled tweets and the *present* or 1 label only 403 times. This may have caused the kappa to underestimate the actual inter-annotator agreement. For the user-level annotation, the proposed scheme achieved an inter-annotator agreement 0.83 or *almost perfect.* Similar to the tweet-level annotation, as presented in Section 7.2.3, *vulnerable to radicalisation* and *already radicalised* represented minority classes. Considering the chance of underestimation, both kappas were considered to be good.

In all, the process proposed in relation to Research Question 2 was considered good. The process filtered out irrelevant accounts, achieved a good inter-annotator agreement and the number of identified users are considered reasonable based on previous studies. However, the process of collecting data and manually verifying tweets is resource-intensive and consists of many steps. Future work should aim to build on the insights of this study to automate the identification phase further.

Research Question 3 *Do people at risk of radicalisation share any common personality traits and do these traits differ from regular users?*

The personality prediction models trained as part of Experiment 1 were aimed at answering Research Question 3. The models were trained and tested on the *personality*$_d$ dataset, before a selection of models were used for predicting the personality traits of users in *radicalisation*$_d$ and *non-radical*$_d$. The first part of Research Question 3 aims at exploring common traits among the users of *radicalisation*$_d$. The second part explores similarities and differences in personality traits between *radicalisation*$_d$ and *non-radical*$_d$. The results of the experiment can be seen in Figure 9.2-9.6.

For Extraversion, as can be seen from Figure 9.2a, there was no clear personality score among the users of *radicalisation$_d$*. The users were normalised around 2.50-4.30. The majority of users were found around 3.40-3.70, however, there was no peak in terms of score. In comparison, the users of *non-radical*, found in Figure 9.2b, were also normalised in the same area. However, the non-radical users had a clear peak around 3.55. The results show that people at risk of radicalisation may not exhibit any clear similarities in terms of Extraversion. They differ somewhat from non-radical users, being that the distribution does not peak.

For Agreeableness, as can be seen from Figure 9.3a, there was some interesting findings for users at risk of radicalisation. The predicted scores for the users of *radicalisation$_d$* did not yield a normal distribution. Instead the users were distributed in a narrow score area of 3.10-3.55, with the majority having a score of 3.25-3.40. The results suggest that users at risk of radicalisation share common characteristics in terms of agreeableness. When comparing the results with those found in Figure 9.3b, the *non-radical$_d$* users are distributed in the range 3.10-3.70. Despite the two domains being within the same area, their distribution differ somewhat. While nearly half of the *radicalisastion$_d$* users had a score of around 3.25 for agreeableness, just over one third of the *non-radical$_d$* users had the same score. However, despite the domains having differences, they both fall inside the same score range, which may suggest that there are certain similarities between the two groups.

For Neuroticism, as can be seen from Figure 9.4a, there were no clear personality score for users at risk of radicalisation. The main part of the scores had a normal distribution around 2.20-3.70, with a small peak around 3.10. However, when comparing these results to that of *non-radical$_d$*, found in Figure 9.4b, the results become interesting. The predictions for *non-radical$_d$* showed only 2% of the users were below a score of 2.80, with the main distribution being within the range 2.80-3-70. In comparison, nearly 10% of the *radicalisation$_d$* users had a score below 10%. The main proportion of *non-radical$_d$* users were more evenly distributed around 2.80-3.25, with a small peak around 3.25. For *radicalisation$_d$* had a clearer peak around 3.10, suggesting that ordinary Twitter users score higher on Neuroticism than users at risk of radicalisation. Though marginal, this finding was surprising as several of the indicators defined had a positive correlation with Neuroticism.

For Openness, the results were interesting. For *radicalisation$_d$*, as can be seen from Figure 9.5a, the scores were distributed within the range 3.70-4.15, with a peak around 3.85, suggesting a medium high score for this domain. For *non-radical$_d$* the distribution was completely different. The distribution fell within the same range as that of *radicalisation$_d$*, but was skewed right, with a peak around 4.15. The results were considered interesting. Despite being above the average score for Openness, *radicalisation$_d$* still scored lower than that of *non-radical$_d$*. These results may be said to contradict the findings of Wiktorowics (2005), suggesting that individuals being radicalised has an openness to new world views. Though scoring in the higher ranges on Openness, the users of *radicalisation* still scored lower than the users of *non-radical$_d$*.

For conscientiousness, as can be seen from Figure 9.6a, the users vulnerable to radicalisation were distributed in the range 3.25-3.80, with a clear peak around 3.40. The results indicate a clear similarity within the domain, having an openness score just above the average. In comparison, *non-radical$_d$* were distributed in the range 3.25-4.00, with no clear peak. The majority of these users were evenly distributed around 3.40-3.55. Though marginal, the users at risk of radicalisation scores lower on conscientiousness when compared to the users of *non-radical$_d$*.

Research Question 4 *Do any indicators of radicalisation correlate more with certain personality traits than others?*

In order to answer Research Question 4, the predicted scores from Experiment 2 and the accumulated indicators of the annotated users from the data collection process were used as part of Experiment 3. The correlation between the occurrence of each users' indicators and their predicted traits was then calculated. The results of Experiment 3 can be seen in Table 9.16.

The results provided some interesting information on the predictive power of each indicator and their relationship with the predicted traits. One interesting finding was that *Frustration* was found to have close to no correlation with the majority of traits, except for Neuroticism. The correlation between Frustration and Neuroticism was in fact the strongest correlation across the indicators and personality traits, with a score of 0.27. *Negative words* as both weak positive and negative correlation across all traits, with the highest being for Openness, with a score of 0.21. *Discrimination* did not yield any substantial correlation with the indicators, but still achieved the highest overall correlation for Extraversion, with a score of 0.16. *Negative West* had a correlation of 0.21 for Neuroticism and 0.23 for Openness. *Pro-Jihad* was the only indicator, though being weak, with negative correlations only.

When viewing the correlations from the personality trait perspective, Neuroticism achieved the highest average absolute correlation of 0.184 across the indicators. This indicated that Neuroticism was to some degree related to all the defined indicators and may suggest that the emotional stability is a prevalent factor during the radicalisation phase. Conscientiousness had to particularly strong correlations, while also yielding the lowest average absolute correlation of 0.068 across the traits. This was considered interesting, as users of *radicalisation$_d$* showed signs of sharing traits with regards to Conscientiousness, but the trait were not seemingly correlated with any of the defined indicators.

In summary, it was found that Frustration correlated more with Neuroticism, Negative Words correlated more with Openness, and Negative West correlated more with both Neuroticism and Openness. Despite being a weak correlation, Discrimination correlated the most of all indicators with regards to Extraversion. The findings indicate that several of the derived indicators from Research Question 1 may in fact have a degree of predictive power for certain personality traits.

Goal *Investigate the personality traits of individuals at risk of radicalisation on social media by training an automatic personality prediction model using linguistic cues.*

On the basis of the experiments conducted as part of this Thesis, the research questions presented in Section 1.2, and the evaluation of these questions, the goal of this Thesis is considered to be thoroughly addressed and answered. A dataset of users vulnerable to radicalisation has been collected, a regression model for predicting personality traits was implemented and the personality traits of $personality_d$ and $non\text{-}radical_d$ has been predicted. The experiments show that individuals at risk of radicalisation share common characteristics in terms of personality traits. The similarities were most clear for the Conscientiousness and Agreeableness traits, showing that the majority of these individuals fall within the same score range. When comparing the results with those obtained for $non\text{-}radical_d$ there are differences in the score distribution, but these differences do not appear to be significant. As this study is the first to conduct personality prediction for people at risk of radicalisation, future research should aim to explore the differences between users vulnerable to radicalisation and ordinary users more. The $radicalisation_d$ and *non-radical* datasets used for the final prediction consisted of only 259 users each. Future research should aim to expand these datasets with more users to validate these differences further.

11. Conclusion and Future Work

Dissemination of radical content and recruitment of individuals to extreme ideologies on social media is still a reality, despite substantial efforts from the platforms themselves. As was evident from this Thesis, a total of 121 radical users were able to obtain close to 73,000 followers, enabling these radical users to severely impact their viewers. The detection of potential recruits is an important step in hindering the growth of radical communities. In an effort to improve the identification of individuals at risk of radicalisation on social media, this Thesis investigates the effects of applying NLP techniques for predicting personality traits of these individuals. As stated in Chapter 1, this Thesis does by no means assume there to be an ultimate profile of people at risk of being radicalised or engaging in radical activities. Forty years of research of radicalism has rejected the notion of there being a specific person that engages in radicalism. Unfortunately, this has resulted in an unfair picture of the potential for psychological contributions to the field (Horgan, 2008). This Thesis sets out to explore personality as a driving factor to radicalism by viewing personality as a series of individual traits, using the Big5 Personality Model. Several regression architectures were tested to derive a set of models capable of carrying out personality prediction with reasonable accuracy. The resulting models were run on a manually collected dataset of users believed to at risk of radicalisation. The study has successfully managed to identify similarities in certain personality traits among these users. The results were compared to a manually collected dataset of what was believed to be regular non-radical users. The results show minor differences in terms of the predicted traits and with regards to the score distribution. Despite interesting results, there are several aspects which needs to be taken into consideration in order to verify the significance of these differences.

This chapter provides a conclusion on the work conducted as part of this Thesis. In addition, the chapter gives an elaboration of the contributions to the field of radicalisation research and suggestions for future work to build upon the results of this Thesis.

11.1. Contributions

This Thesis contributes to the field of radicalisation research on social media by shifting the focus from understanding the characteristics of individuals already radicalised, to those at risk of being drawn into radical communities. There is a limited amount of research on people prior to being radicalised on social media. During the literature review, only the study by Rowe and Saif (2016) was found to harvest the potential of

NLP for identifying characteristics of individuals prior to radicalisation.

A literature review focusing on previous studies on pathways and factors leading to radicalisation resulted in a set of indicators for identifying people vulnerable to radicalisation. Four new indicators were introduced for identifying and annotating these users and resulted in a final dataset of 259 users, referred to as $radicalisation_d$. The inter-annotator agreement proved the defined data collection methodology and annotation scheme to be reliable, yielding a dataset and data collection methodology that can serve as a foundation for future work. $radicalisation_d$ is considered to be one of the main contributions of this Thesis and possibly one of the first manually annotated datasets focusing purely on social media users prior to radicalisation.

The experiments conducted as part of this Thesis resulted in several contributions. This Thesis contributes with an expansion of the radicalisation research field by utilising state-of-the-art methods of personality prediction within a new domain. The results of the experiments provide new insight into the traits of individuals considered to be at risk of radicalisation on social media by showing certain similarities in terms of personality profile among the users identified. The experiments also show that the personality of Twitter users considered to be at risk of radicalisation differ to some degree to that of ordinary Twitter users. This new insight can in turn provide useful information for how to best target individuals on social media prior to them being radicalised. The experiments also provide information on the predictive power of indicators affiliated with radicalisation by measuring their correlation with predicted personality traits. The results of the experiments show that several of the defined indicators to a certain degree correlate with both the Neuroticism and Openness part of an individual's personality.

11.2. Future Work

During the process of working with this Thesis, several ideas for improvement arose. The following section provides information on potential future work to build upon the results and findings of this Thesis.

11.2.1. Expanding Dictionary

For the purpose of collecting data on users vulnerable to radicalisation and filtering out irrelevant accounts, a dictionary of common terms associated with radicalisation was comprised. As part of analysing the resulting dataset, a word cloud of the most common terms among the identified users was generated. While several of the terms identified in the word cloud could be categorised according to the defined indicators, the majority of the identified terms were not contained in the dictionary. As part of future work, the newly identified terms should be incorporated in the dictionary to see if these terms better help identify relevant users and reduce the number of users annotated with the *unrelated* label.

11.2.2. Expanding Radicalisation Dataset

As part of evaluating the research goal in Section 10.2, questions about the size of *radicalisation*$_d$ and *non-radical*$_d$ were raised. While the final prediction conducted as part of experiment 2 resulted in interesting insight, comparison between the two domains may have been subject to sparse amounts of data. The prediction was deemed reliable, but future work should aim to expand the datasets of users vulnerable to radicalisation and ordinary users to explore potential similarities and differences further.

11.2.3. Datasets for Personality Prediction

As discussed in Section 10.1.3, the quality of a trained model is only as good as the data it is trained on. For this Thesis, the *personality*$_d$ was comprised on three datasets and used for training models for personality prediction. Chapter 7 presented an overview of the distribution of the personality scores for each trait contained in the myPersonality Twitter dataset and PAN2015 Author Profiling dataset. From the distribution it was clear that there was a lack of instances for the lower and upper bounds of the scores, resulting in few examples for the models to train on. As part of future work, one should aim to extend the existing datasets with more examples for other value ranges within the traits. This may provide new insight into the personality traits of the users contained in *radicalisation*$_d$ and may provide a new picture on the distribution of traits.

Another topic for discussion was the conversion from binary to numerical values for the stream-of-conscientiousness dataset. While this was considered a promising approach based on the results of Nordnes and Gran (2019), future research should investigate the results of excluding this dataset from the training process. As previously discussed in Chapter 10, the use of average numerical values as a representation for the binary values may introduce overfitting to the models.

The sparsity of the *personality*$_d$ may also have impacted the result of the trained models. Despite the results of the prediction models being considered reliable, future research should aim to extend the size of the training dataset to further increase the accuracy of the trained models.

11.2.4. Alternative Personality Models

Based on the reviewed research, this Thesis chose to focus on the Big5 personality model, being both a reliable and recognised model. It is however important to emphasise that an individual's personality consists of several aspects which the Big5 personality model may not successfully capture. Future research should aim to explore other personality models in an effort to detect new traits that can be used for identifying characteristics of individuals at risk of radicalisation. The Dark Triad has contributed to interesting insight into the habits of individuals spreading hateful content on social media (Sumner et al., 2012; Preotiuc-Pietro et al., 2016). Exploring the darker traits of individuals at risk of radicalisation may provide additional insight that can be used for identifying these individuals on social media.

11.2.5. Personality Traits for Identification

As part of the experiments conducted in this Thesis, certain similarities in terms of personality traits were found for users considered to be at risk of radicalisation. For several of these traits, the score distribution differed somewhat from the users of *non-radical*$_d$. As part of identifying individuals at risk of radicalisation, future work should explore the effects of incorporating personality as a feature for automatic detection of these users on social media.

Bibliography

Swati Agarwal and Ashish Sureka. Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter. In Raja Natarajan, Gautam Barua, and Manas Ranjan Patra, editors, *Distributed Computing and Internet Technology*, Lecture Notes in Computer Science, pages 431–442. Springer International Publishing, Cham, Switzerland, 2015. ISBN 9783319149769. URL https://doi.org/10.1007/978-3-319-14977-6_47.

Jay Alammar. The illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning. *Blog Post*, December 2018. URL http://jalammar.github.io/illustrated-bert/.

Oscar Araque and Carlos A Iglesias. An approach for radicalization detection based on emotion signals and semantic similarity. *IEEE Access*, 8:17877–17891, 2020. URL https://doi.org/10.1109/ACCESS.2020.2967219.

Oscar Araque and Carlos A. Iglesias. An Ensemble Method for Radicalization and Hate Speech Detection Online Empowered by Sentic Computing. *Cognitive Computation*, 14(1):48–61, January 2022. URL https://doi.org/10.1007/s12559-021-09845-6.

Pierre-Hadrien Arnoux, Anbang Xu, Neil Boyette, Jalal Mahmud, Rama Akkiraju, and Vibha Sinha. 25 Tweets to Know You: A New Model to Predict Personality with Social Media. *ArXiv*, April 2017. URL https://arxiv.org/ftp/arxiv/papers/1704/1704.05513.pdf.

Muhammad Asif, Atiab Ishtiaq, Haseeb Ahmad, Hanan Aljuaid, and Jalal Shah. Sentiment analysis of extremism in social media from textual information. *Telematics and informatics*, 48:101345, May 2020. ISSN 0736-5853. URL https://doi.org/10.1016/j.tele.2020.101345.

Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and patterns of Facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 24–32, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312288. URL https://doi.org/10.1145/2380718.2380722.

Thomas Bayes and Richard Price. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53: 370–418, 1763. ISSN 02607085. URL http://www.jstor.org/stable/105741.

Matthew C. Benigni, Kenneth Joseph, and Kathleen M. Carley. Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PloS one*, 12(12):e0181405, December 2017. ISSN 1932-6203. URL `https://doi.org/10.1371/journal.pone.0181405`.

Elizabeth Bodine-Baron, Todd C Helmus, Madeline Magnuson, and Zev Winkelman. *Examining ISIS support and opposition networks on Twitter*. RAND, Santa Monica, CA, October 2016. ISBN 978-0-8330-9589-3. URL `https://www.rand.org/pubs/research_reports/RR1328.html`.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, June 2017. ISSN 2307-387X. URL `https://doi.org/10.1162/tacl_a_00051`.

Randy Borum. Understanding the Terrorist Mindset. *FBI Law Enforcement Bulletin*, 72, March 2003. URL `https://www.ojp.gov/pdffiles1/nij/grants/201462.pdf`.

Randy Borum. Radicalization into Violent Extremism II: A Review of Conceptual Models and Empirical Research. *Journal of Strategic Security*, 4(4):37–62, December 2011a. URL `https://doi.org/10.5038/1944-0472.4.4.2`.

Randy Borum. Radicalization into Violent Extremism I: A Review of Social Science Theories. *Journal of Strategic Security*, 4(4):7–36, December 2011b. URL `https://doi.org/10.5038/1944-0472.4.4.1`.

Giulio Carducci, Giuseppe Rizzo, Diego Monti, Enrico Palumbo, and Maurizio Morisio. TwitPersonality: Computing Personality Traits from Tweets Using Word Embeddings and Supervised Learning. *Information (Switzerland)*, 9(5), May 2018. URL `http://doi.org/10.3390/info9050127`.

Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. URL `https://doi.org/10.1177/001316446002000104`.

Katie Cohen and Lisa Kaati. Digital Jihad: Propaganda from the Islamic State. *Swedish Defence Research Agency (FOI)*, November 2018. ISSN 1650-1942. URL `https://www.foi.se/report-summary?reportNo=FOI-R--4645--SE`.

Maura Conway, Moign Khawaja, Suraj Lakhani, Jeremy Reffin, Andrew Robertson, and David Weir. Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts. *Studies in Conflict & Terrorism*, 42(1-2):141–160, 2019. URL `https://doi.org/10.1080/1057610X.2018.1513984`.

Clayton Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. BotOrNot: A System to Evaluate Social Bots. *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, February 2016. URL `https://doi.org/10.1145/2872518.2889302`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL `https://doi.org/10.18653/v1/N19-1423`.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26 (3):297–302, July 1945. URL `https://doi.org/10.2307/1932409`.

Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. Recognising personality traits using Facebook status updates. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(2):14–18, August 2021. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14470`.

Miriam Fernandez, Moizzah Asif, and Harith Alani. Understanding the Roots of Radicalisation on Twitter. In *Proceedings of the 10th ACM Conference on web science*, WebSci '18, pages 1–10, New York, NY, USA, May 2018. Association for Computing Machinery. ISBN 9781450355636. URL `https://doi.org/10.1145/3201064.3201082`.

Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. Predicting Online Extremism, Content Adopters, and Interaction Reciprocity. In Emma Spiro and Yong-Yeol Ahn, editors, *Social Informatics*, Lecture Notes in Computer Science, pages 22–39, Cham, 2016. Springer International Publishing. ISBN 9783319478739. URL `http://dx.doi.org/10.1007/978-3-319-47874-6_3`.

John Rupert Firth. *A Synopsis of Linguistic Theory, 1930-1955.* Studies in Linguistic Analysis (special volume of the Philological Society). The Philological Society, Oxford, 1957.

Ali Fisher. Swarmcast: How Jihadist Networks Maintain a Persistent Online Presence. *Perspectives on terrorism (Lowell)*, 9(3):3–20, June 2015. ISSN 23343745. URL `http://www.jstor.org/stable/26297378`.

Donald W. Fiske. Consistency of the factorial structures of personality ratings from different sources. *Journal of abnormal and social psychology*, 44(3):329–344, July 1949. ISSN 0096-851X. URL `https://doi.org/10.1037/h0057198`.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions.* Wiley Series in Probability and Statistics. John Wiley & Sons, Nashville, Tennessee, 3 edition, September 2003. ISBN 0471526290.

Daveed Gartenstein-Ross, Joshua D. Goodman, and Laura Grossman. *Terrorism in the West 2008: A Guide to Terrorism Events and Landmark Cases.* Foundation for Defense of Democracies, August 2009. ISBN 0981971229.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting Personality from Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156. IEEE, October 2011a. ISBN 9781457719318. URL https://doi.org/10.1109/PASSAT/SocialCom.2011.33.

Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *CHI '11 Extended Abstracts on human factors in computing systems*, CHI EA '11, pages 253–262, New York, NY, USA, January 2011b. Association for Computing Machinery. ISBN 9781450302685. URL https://doi.org/10.1145/1979742.1979614.

Lewis R. Goldberg. An Alternative "Description of Personality": The Big-Five Factor Structure. *Journal of personality and social psychology*, 59(6):1216–1229, December 1990. ISSN 0022-3514. URL https://doi.org/10.1037//0022-3514.59.6.1216.

Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, August 1954. URL https://doi.org/10.1080/00437956.1954.11659520.

John Horgan. From Profiles to Pathways and Roots to Routes: Perspectives from Psychology on Radicalization into Terrorism. *The ANNALS of the American Academy of Political and Social Science*, 618(1):80–94, July 2008. URL https://doi.org/10.1177/0002716208317539.

Michael Jensen, Patrick James, Gary LaFree, Aaron Safer-Lichtenstein, and Elizabeth Yates. The Use of Social Media by United States Extremists. START, 2018. URL www.start.umd.edu/pubs/START_PIRUS_UseOfSocialMediaByUSExtremists_ResearchBrief_July2018.pdf.

Jonathan Kenyon, Jens Binder, and Christopher Baker-Beall. Exploring the role of the Internet in radicalisation and offending of convicted extremists. Ministry of Justice Analytical Series, pages 10–11. Ministry of Justice, London, UK, 2021. ISBN 9781840999761. URL https://www.gov.uk/government/publications/exploring-the-role-of-the-internet-in-radicalisation-and-offending-of-convicted-extremists.

Anders Kofod-Pedersen. How to do a Structured Literature review in Computer Science. Technical report, Technical report, Department of Computer Science, Norwegian University of Science and Technology Science, Trondheim, Norway, 2018.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. URL https://doi.org/10.18653/v1/D18-2012.

Pavan K. N. Kumar and Marina L. Gavrilova. Personality Traits Classification on Twitter. In *2019 16th IEEE International Conference on Advanced Video and Signal*

Based Surveillance (AVSS), pages 1–8, Taipei, Taiwan, September 2019. IEEE. URL `https://doi.org/10.1109/AVSS.2019.8909839`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*, September 2020. URL `https://openreview.net/forum?id=H1eA7AEtvS`.

J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL `https://doi.org/10.2307/2529310`.

Raúl Lara-Cabrera, Antonio González Pardo, Karim Benouaret, Noura Faci, Djamal Benslimane, and David Camacho. Measuring the Radicalisation Risk in Social Networks. *IEEE Access*, 5:10892–10900, 2017. URL `https://doi.org/10.1109/ACCESS.2017.2706018`.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. Data Mining. In *Mining of Massive Datasets*, pages 1–17. Cambridge University Press, Cambridge, England, 3 edition, January 2020. ISBN 9781108476348.

Ana Carolina E.S. Lima and Leandro Nunes de Castro. A multi-label, semi-supervised classification approach applied to personality prediction in social media. *Neural networks*, 58:122–130, October 2014. ISSN 0893-6080. URL `https://doi.org/10.1016/j.neunet.2014.05.020`.

Fei Liu, Julien Perez, and Scott Nowson. A Language-independent and Compositional Model for Personality Trait Recognition from Short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 754–764, Stroudsburg, PA, USA, April 2017. Association for Computational Linguistics. URL `https://doi.org/10.18653/v1/E17-1071`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, July 2019. URL `https://arxiv.org/abs/1907.11692`.

Ragnhild Lygre, Jarle Eid, Gerry Larsson, and Magnus Ranstorp. Terrorism as a process: A critical review of Moghaddam's "Staircase to Terrorism". *Scandinavian journal of psychology*, 52(6):609–16, September 2011. URL `https://doi.org/10.1111/j.1467-9450.2011.00918.x`.

Chenglong Ma, Weiqun Xu, Peijia Li, and Yonghong Yan. Distributional Representations of Words for Short Text Classification. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 33–38. Association for Computational Linguistics, January 2015. URL `https://doi.org/10.3115/v1/W15-1505`.

Valeria Maeda-Gutiérrez, Carlos E Galván-Tejada, Laura A Zanella-Calzada, José M Celaya-Padilla, Jorge I Galván-Tejada, Hamurabi Gamboa-Rosales, Huizilopoztli Luna-García, Rafael Magallanes-Quintanar, Carlos A Guerrero Méndez, and Carlos A Olvera-Olvera. Comparison of convolutional neural network architectures for classification of tomato plant diseases. *Applied Sciences (Basel)*, 10(4):1245, February 2020. URL `https://doi.org/10.3390/app10041245`.

Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE intelligent systems*, 32(2):74–79, March 2017. ISSN 1541-1672. URL `https://doi.org/10.1109/MIS.2017.23`.

Andrey A. Markov. An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Science in Context*, 19(4):591–600, December 2006. URL `https://doi.org/10.1017/S0269889706001074`.

Robert R. McCrae and Oliver P. John. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2):175–215, June 1992. ISSN 0022-3506. URL `https://doi.org/10.1111/j.1467-6494.1992.tb00970.x`.

Jobo W. McHoskey, William P. Worzel, and Chris Szyarto. Machiavellianism and psychopathy. *Journal of personality and social psychology*, 74(1):192–210, January 1998. URL `https://doi.org/10.1037//0022-3514.74.1.192`.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. Bottom-Up and Top-Down: Predicting Personality with Psycholinguistic and Language Model Features. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1184–1189. IEEE, November 2020. ISBN 1728183162. URL `https://doi.org/10.1109/ICDM50108.2020.00146`.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at ICLR*, pages 1–12, January 2013. URL `https://arxiv.org/abs/1301.3781`.

Thom M. Mitchell. *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill Professional, New York, NY, USA, 1 edition, March 1997. ISBN 0070428077.

Fathali Moghaddam. The Staircase to Terrorism: A Psychological Exploration. *The American psychologist*, 60(2):161–169, February 2005. URL `https://doi.org/10.1037/0003-066X.60.2.161`.

Isabel B. Myers. *MBTI Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator*. Consulting Psychologists Press, 3 edition, January 1998. ISBN 9780891061304.

Andrea Hollung Nordnes and Martine Alvilde Gran. Automatic Classification of Pro-Eating Disorder Twitter Accounts with Personality as a Feature. MSc Thesis, Dept.

of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway, May 2019.

Derek O'Callaghan, Nico Prucha, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. Online social media in the Syria conflict: Encompassing the extremes and the in-betweens. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 409–416. IEEE, August 2014. URL `https://doi.org/10.1109/ASONAM.2014.6921619`.

Keiron O'Shea and Ryan Nash. An Introduction to Convolutional Neural Networks. November 2015. URL `https://arxiv.org/abs/1511.08458`.

D. Parekh, Amarnath Amarasingam, Lorne Dawson, and D. Ruths. Studying jihadists on social media: A critique of data collection methodologies. 12(3):3–21, June 2018. ISSN 2334-3745.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. Automatic Personality Assessment Through Social Media Language. *Journal of personality and social psychology*, 108(6):934–952, November 2015. ISSN 0022-3514. URL `https://doi.org/10.1037/pspp0000020`.

Karl Pearson. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society London*, 58(347-352):240–242, December 1895. URL `https://doi.org/10.1098/rspl.1895.0041`.

Kuei-Hsiang Peng, Li-Heng Liuo, Cheng-Shang Chang, and Duan-Shin Lee. Predicting personality traits of chinese users based on Facebook wall posts. In *2015 24th Wireless and Optical Communication Conference (WOCC)*, pages 9–14. IEEE, October 2015. URL `https://doi.org/10.1109/WOCC.2015.7346106`.

James Pennebaker and Laura King. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296–312, December 1999. URL `https://doi.org/10.1037/0022-3514.77.6.1296`.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Stroudsburg, PA, USA, October 2014. Association for Computational Linguistics. URL `https://doi.org/10.3115/v1/D14-1162`.

Barbara Plank and Dirk Hovy. Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98. Association for Computational Linguistics, September 2015. URL `https://doi.org/10.18653/v1/W15-2913`.

Tomas Precht. Home grown terrorism and Islamist radicalisation in Europe: From conversion to terrorism. December 2007. URL https://www.justitsministeriet.dk/sites/default/files/media/Arbejdsomraader/Forskning/Forskningspuljen/2011/2007/Home_grown_terrorism_and_Islamist_radicalisation_in_Europe_-_an_assessment_of_influencing_factors__2_.pdf.

Daniel Preotiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. Studying the Dark Triad of Personality through Twitter Behavior. In *Proceedings of the 25th ACM International on conference on information and knowledge management*, CIKM '16, pages 761–770, New York, NY, USA, October 2016. Association for Computing Machinery. ISBN 9781450340731. URL https://doi.org/10.1145/2983323.2983822.

Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 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*, pages 180–185. IEEE, October 2011. ISBN 9781457719318. URL https://doi.org/10.1109/PASSAT/SocialCom.2011.26.

Carl Edward Rasmussen and Christopher K I Williams. *Gaussian processes for machine learning*. Adaptive Computation and Machine Learning series. MIT Press, London, England, November 2005. ISBN 026218253X.

Shahamak Rezaei and Marco Goli. *House of War: Islamic Radicalisation in Denmark*. Centre for Studies in Islamism and Radicalization (CIR), Aarhus University, Denmark, January 2010. ISBN 9788792540089.

Brooke Rogers and Peter Neumann. Recruitment and Mobilisation for the Islamist Militant Movement in Europe. European Commission, January 2007. URL https://icsr.info/2008/10/01/recruitment-and-mobilisation-for-the-islamist-militant-movement-in-europe/.

Matthew Rowe and Hassan Saif. Mining Pro-ISIS Radicalisation Signals from Social Media Users. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), March 2016. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14716.

Stuart Russell and Peter Norvig. *Artificial Intelligence*. Pearson, Upper Saddle River, NJ, 3 edition, December 2009. ISBN 0136042597.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, March 2020. URL https://arxiv.org/abs/1910.01108.

Mike Schuster and Kaisuke Nakajima. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE, March 2012. URL https://doi.org/10.1109/ICASSP.2012.6289079.

Andrew H. Schwartz, Johannes C. Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PloS one*, 8(9):e73791–e73791, 2013. ISSN 1932-6203. URL https://doi.org/10.1371/journal.pone.0073791.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. URL https://doi.org/10.18653/v1/P16-1162.

Mitchell Silber and Arvin Bhatt. *Radicalization in the West: The Homegrown Threat: The NYPD Jihadist Report.* OccupyBawlStreet.com Press, January 2015. ISBN 0692371702.

Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An Energy-Efficient, High-Performance GPGPU Computing Research Infrastructure, 2021. URL https://arxiv.org/abs/1912.05848.

Marieke Slootman, F. Demant, F. Buijs, and Jean Tillie. *Processes of Radicalisation. Why some Amsterdam Muslims become radicals.* IMES, Amsterdam, Netherlands, October 2006.

Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J. Park. Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 386–393. IEEE, 2012. URL https://doi.org/10.1109/ICMLA.2012.218.

Thorvald Sørensen. Method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Bioliske Skrifter*, 5, January 1948.

Tommy Tandera, Hendro, Derwin Suhartono, Rini Wongso, and Yen Lina Prasetio. Personality Prediction System from Facebook Users. *Procedia computer science*, 116: 604–611, December 2017. ISSN 1877-0509. URL https://doi.org/10.1016/j.procs.2017.10.016.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, March 2010. URL https://doi.org/10.1177/0261927X09351676.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *ArXiv*, June 2017. URL `http://arxiv.org/abs/1706.03762`.

Quintan Wiktorowics. *Radical Islam rising: Muslim extremism in the West*. Rowman & Littlefield, Inc, Lanham, Maryland, July 2005. ISBN 9781461641711.

William E. Winkler. Data Cleaning Methods. Washington, DC, USA, November 2003. U.S. Bureau of the Census Statistical Research. URL `https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.2066&rep=rep1&type=pdf`.

Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. Deep learning-based personality recognition from text posts of online social networks. *Applied intelligence (Dordrecht, Netherlands)*, 48(11):4232–4246, November 2018. ISSN 0924-669X. URL `https://doi.org/10.1007/s10489-018-1212-4`.

Appendices

A. Primary and Secondary Inclusion Criteria

Separate inclusion criteria were formulated for the two query strings. These criteria were used as part of the Structured Literature Review to select primary studies and filter out irrelevant papers.

A.1. Primary Inclusion Criteria - 1st Query

IC1: The study's main concern is predicting people turning to extremism

IC2: The study is a primary study presenting empirical results

A.2. Secondary Inclusion Criteria - 1st Query

IC3: The study focuses on predicting people turning to extremism based on written data from Twitter or Facebook

IC4: The study describes an implementation of an algorithm for predicting people turning to extremism

A.3. Primary Inclusion Criteria - 2nd Query

IC1: The study's main concern is prediction of personality

IC2: The study is a primary study presenting empirical results

A.4. Secondary Inclusion Criteria - 2nd Query

IC3: The study focuses on predicting personality based on written data from Twitter and Facebook

IC4: The study describes an implementation of an algorithm for predicting personality

B. Quality Assessment Criteria

Quality assessment criteria were used for papers passing the secondary inclusion criteria, as part of the Structured Literature Review. Papers were awarded points based on quality criteria QC1-QC10. For each of the criteria, 1 point was given for full fulfilment, 1/2 point for partially fulfilled, and 0 points if the criteria were not met. The quality assessment scores of the papers were used as a metric to select the final papers for this Thesis.

QC1: Is there a clear statement of the aim of the study?

QC2: Is the study put into the context of other studies and research?

QC3: Are system or algorithmic design decisions justified?

QC4: Is the test data set reproducible?

QC5: Is the study algorithm reproducible?

QC6: Is the experimental procedure thoroughly explained?

QC7: Is it clearly stated in the study which other algorithms the study's algorithm(s) have been compared with?

QC8: Are the performance metrics used in the study clearly explained?

QC9: Are the test results thoroughly analyzed?

QC10: Does the test evidence support the findings presented?

C. Extracted Data for Automatic Personality Prediction

The following data was extracted from the initial literature review on automatic personality prediction.

- Unique ID
- Authors
- Title
- Publication year
- Personality Model

- Algorithm used for personality prediction
- Selected features
- Dataset
- Relevant finding and conclusion

D. Extracted Data for Identification of People Vulnerable to Social Media Extremism

The following data was extracted from the initial literature review on identification of people vulnerable to social media extremism.

- Unique ID
- Authors
- Title
- Publication year
- Algorithm used for identification of people vulnerable to social media extremism
- Selected features
- Dataset
- Relevant finding and conclusion

E. Quality Assessment Results

Table E.1.: Quality Assessment Scores From the Initial Literature Review

ID	QC1	QC2	QC3	QC4	QC5	QC6	QC7	QC8	QC9	QC10	Score
1	1/2	1	1/2	1/2	1	1	0	1	1/2	1	7
2	1	1	1/2	0	1/2	1	1	0	1	1	7
3	1	1	1	1	1/2	0	1	1/2	1	1	8
4	1	1	1/2	1/2	1/2	1	0	1	1	1	7.5
5	1	1	0	1/2	1/2	1/2	1/2	1/2	1	1/2	6
6	1	1	1	0	1/2	1/2	1	0	1	1	7
7	1	1	0	0	0	0	0	0	1	1	4
8	1	1	1/2	1/2	1/2	1	1/2	1/2	1	1	7.5
9	1	1	0	1/2	1	1/2	0	1/2	1	1	6.5
10	1	1	1	1	1	1/2	1	1/2	1	1	9
11	1	1	0	0	0	1/2	0	0	1	1	4.5
12	1	1	0	0	1/2	1/2	0	1	1/2	1	5.5
13	1	1	1	0	1	1	1/2	1/2	1	1	8
14	1	1	0	0	0	0	0	0	1	1/2	3.5
15	1	1	0	1/2	1/2	1/2	1/2	1	1	1	7
16	1	1	1	1/2	1	1	1	1	1	1	9.5
17	1	1	1/2	1/2	0	1/2	1/2	0	1/2	1	5.5
18	1	1	1	1/2	1/2	1	1/2	1	1	1	8.5
19	1	1	1/2	1/2	1/2	1/2	0	1/2	1	1	6.5
20	1	1	1	1/2	1/2	1/2	1/2	1/2	1	1	7.5
21	1	1	1/2	0	1/2	1	0	1	1	1	7
22	1	1	1/2	1	1/2	1/2	0	1/2	1/2	1	6.5
23	1	1	1	1/2	1/2	1	1/2	1/2	1	1	8
24	1	1	1/2	1	1	1/2	1/2	1/2	1	1	8
25	1	1	1	1/2	0	1	0	1	1	1	7.5

F. Abbreviations

Table F.1.: Abbreviated Words Replaced as Part of Pre-Processing

Abbreviated form	**Full form**
dnt	don't
knw	know
nt	not
rly	really
idk	i dont know
ppl	people
pls	please
plz	please
ty	thank you
tyvm	thank you very much
nvm	nevermind
abt	about
lmk	let me know
fr	for real
rn	right now

G. Dictionary of Terms Affiliated With Radicalisation

G.1. Frustration/Grievance

- balec
- yomb
- shit
- crap
- damn
- fuck
- african attack
- pylytheist
- polytheism
- eurafrica
- intellectual terrorism
- intellectual terrorist
- russian rat
- terrorist leftist

G.2. Negative Words / Hate speech

- carba karab
- cheh
- zamel
- hate
- guilt
- shame
- fault
- faggot
- dirty jew
- white christian

G.3. Negative Ideas About Western Society

- babtou
- chmeta
- guèouri
- jahiliya
- sahawat
- wala wal-bara
- al-walâ-u wa l-brâ
- west
- usa
- impure
- infidel
- hypocrite
- disbeliever
- al-adu al-baid
- al-'adu al bai'd
- oppressors

G.4. Pro-Jihad Terms

- amiliya amniya
- amn Khariji
- ansar
- bid'a
- caliphate
- shahid
- dabiq
- dogma

- ghulat
- hujjaj
- inghimasiyoun
- mu’askarat
- muhajirine
- muhajiroun
- muhâdjirîn
- muhâdjirûn
- mujahid
- moudjahid
- mujahidin
- murjites
- mujahed
- martyr
- ribat
- takfir
- islamic state
- state of islam
- caliphate
- daesh
- mujahideen
- crusader
- jihad
- istishad
- khilafah
- abu musab az-zarqawi
- al-adu al-qarib
- abu mus’ab az-zarqawi
- al-’adu al-qarib

G.5. Theological Terms/Theological References

- adâb
- adab
- èdèb
- edeb
- akhi
- dar al islam
- dawa
- hijrah
- hijra
- ikhwan
- kuffar
- kafir
- kufr
- fāhishah,
- shirk
- apostate

H. Hyperparameters for each Top-Ranking Feature-Model Combination

H.1. With Stopwords Included

LIWC + SVM (Linear Kernel)
Ext: 'C': 5, 'epsilon': 0.2
Agr: 'C': 20, 'epsilon': 0.8
Neu: 'C': 20, 'epsilon': 0.1
Opn: 'C': 15, 'epsilon': 0.2
Con: 'C': 0.1, 'epsilon': 0.8

LIWC + SVM (Non-Linear Kernels)
Ext: 'C': 20, 'epsilon': 0.5, 'kernel': 'rbf'
Agr: 'C': 20, 'epsilon': 0.8, 'kernel': 'rbf'
Neu: 'C': 20, 'epsilon': 0.5, 'kernel': 'rbf'
Opn: 'C': 20, 'epsilon': 0.2, 'kernel': 'rbf'
Con: 'C': 20, 'epsilon': 0.8, 'kernel': 'rbf'

DistilBert (Layer 2) + MLP
Ext: 'activation': 'tanh', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.01
Agr: 'activation': 'tanh', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.005
Neu: 'activation': 'logistic', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.005
Opn: 'activation': 'logistic', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.01
Con: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.01

DistilBert (Layer 5) + MLP
Ext: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.001
Agr: 'activation': 'relu', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.001
Neu: 'activation': 'logistic', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.001
Opn: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.005
Con: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.01

DistilBert (Layer 6) + MLP
Ext: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.001
Agr: 'activation': 'tanh', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.005
Neu: 'activation': 'logistic', 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.001
Opn: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.001
Con: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.01

ALBERT (Layer 2) + MLP
Ext: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.01
Agr: 'activation': 'tanh', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.01
Neu: 'activation': 'logistic', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.01

Opn: 'activation': 'logistic', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.01
Con: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.01

ALBERT (Layer 11) + MLP
Ext: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.01
Agr: 'activation': 'logistic', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.001
Neu: 'activation': 'relu', 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.01
Opn: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.005
Con: 'activation': 'relu', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.005

ALBERT (Layer 12) + MLP
Ext: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.01
Agr: 'activation': 'logistic', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.001
Neu: 'activation': 'tanh', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.005
Opn: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.005
Con: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.01

DistilBert (Layer 2) + SVM
Ext: 'C': 1, 'epsilon': 0.5, 'kernel': 'rbf'
Agr: 'C': 5, 'epsilon': 0.2, 'kernel': 'rbf'
Neu: 'C': 1, 'epsilon': 0.5, 'kernel': 'rbf'
Opn: 'C': 5, 'epsilon': 0.2, 'kernel': 'rbf'
Con: 'C': 0.1, 'epsilon': 0.1, 'kernel': 'linear'

DistilBert (Layer 5) + SVM
Ext: 'C': 10, 'epsilon': 0.2, 'kernel': 'rbf'
Agr: 'C': 5, 'epsilon': 0.1, 'kernel': 'rbf'
Neu: 'C': 5, 'epsilon': 0.5, 'kernel': 'sigmoid'
Opn: 'C': 5, 'epsilon': 0.2, 'kernel': 'rbf'
Con: 'C': 0.1, 'epsilon': 0.5, 'kernel': 'rbf'

DistilBert (Layer 6) + SVM
Ext: 'C': 0.1, 'epsilon': 0.5, 'kernel': 'linear'
Agr: 'C': 1, 'epsilon': 0.2, 'kernel': 'rbf'
Neu: 'C': 0.5, 'epsilon': 0.5, 'kernel': 'rbf'
Opn: 'C': 1, 'epsilon': 0.2, 'kernel': 'rbf'
Con: 'C': 0.5, 'epsilon': 0.2, 'kernel': 'rbf'

ALBERT (Layer 2) + SVM
Ext: 'C': 1, 'epsilon': 0.5, 'kernel': 'rbf'
Agr: 'C': 20, 'epsilon': 0.2, 'kernel': 'rbf'
Neu: 'C': 1, 'epsilon': 0.5, 'kernel': 'rbf'
Opn: 'C': 5, 'epsilon': 0.2, 'kernel': 'rbf'

Con: ’C’: 0.1, ’epsilon’: 0.2, ’kernel’: ’linear’

ALBERT (Layer 11) + SVM
Ext: ’C’: 1, ’epsilon’: 0.5, ’kernel’: ’rbf’
Agr: ’C’: 10, ’epsilon’: 0.2, ’kernel’: ’rbf’
Neu: ’C’: 1, ’epsilon’: 0.5, ’kernel’: ’rbf’
Opn: ’C’: 1, ’epsilon’: 0.2, ’kernel’: ’rbf’
Con: ’C’: 0.1, ’epsilon’: 0.5, ’kernel’: ’rbf’

ALBERT (Layer 12) + SVM
Ext: ’C’: 0.5, ’epsilon’: 0.5, ’kernel’: ’rbf’
Agr: ’C’: 10, ’epsilon’: 0.2, ’kernel’: ’rbf’
Neu: ’C’: 5, ’epsilon’: 0.5, ’kernel’: ’rbf’
Opn: ’C’: 1, ’epsilon’: 0.2, ’kernel’: ’rbf’
Con: ’C’: 0.1, ’epsilon’: 0.5, ’kernel’: ’rbf’

DistilBert (Layer 2) + GP
Ext: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)

DistilBert (Layer 5) + GP
Ext: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)

DistilBert (Layer 6) + GP
Ext: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)

ALBERT (Layer 2) + GP
Ext: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: ’kernel’: RBF(length_scale=1) + WhiteKernel(noise_level=1)

ALBERT (Layer 11) + GP
Ext: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)

ALBERT (Layer 12) + GP
Ext: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)

DistilBert (Layer 2) + AdaBoost
Ext: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 200
Agr: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Neu: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 100
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Con: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 150

DistilBert (Layer 5) + AdaBoost
Ext: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 100
Agr: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 300
Neu: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 150
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 300
Con: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 100

DistilBert (Layer 6) + AdaBoost
Ext: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 50
Agr: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 50
Neu: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 200
Con: 'learning_rate': 0.01, 'loss': 'square', 'n_estimators': 100

ALBERT (Layer 2) + AdaBoost
Ext: 'learning_rate': 0.01, 'loss': 'square', 'n_estimators': 50
Agr: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 200
Neu: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Con: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 100

ALBERT (Layer 11) + AdaBoost
Ext: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 50
Agr: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 200
Neu: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 200
Opn: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 200
Con: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 100

ALBERT (Layer 12) + AdaBoost
Ext: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 100
Agr: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 200
Neu: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 100
Opn: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 300
Con: 'learning_rate': 0.01, 'loss': 'square', 'n_estimators': 50

H.2. With Stopwords Removed

TF-IDF + SVM
Ext: 'C': 5, 'epsilon': 0.1, 'kernel': 'rbf'
Agr: 'C': 10, 'epsilon': 0.2, 'kernel': 'rbf'
Neu: 'C': 0.5, 'epsilon': 0.2, 'kernel': 'rbf'
Opn: 'C': 10, 'epsilon': 0.1, 'kernel': 'rbf'
Con: 'C': 0.5, 'epsilon': 0.2, 'kernel': 'rbf'

DistilBert (Layer 2) + MLP
Ext: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.001
Agr: 'activation': 'logistic', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.01
Neu: 'activation': 'tanh', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.005
Opn: 'activation': 'tanh', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.005
Con: 'activation': 'logistic', 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.001

DistilBert (Layer 5) + MLP
Ext: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.001
Agr: 'activation': 'relu', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.005
Neu: 'activation': 'tanh', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.001
Opn: 'activation': 'logistic', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.001
Con: 'activation': 'logistic', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.005

DistilBert (Layer 6) + MLP
Ext: 'activation': 'logistic', 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.001
Agr: 'activation': 'relu', 'hidden_layer_sizes': (100,), 'learning_rate_init': 0.05
Neu: 'activation': 'logistic', 'hidden_layer_sizes': (500,), 'learning_rate_init': 0.001
Opn: 'activation': 'logistic', 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.001
Con: 'activation': 'logistic', 'hidden_layer_sizes': (200,), 'learning_rate_init': 0.01

ALBERT (Layer 2) + MLP
Ext: ’activation’: ’logistic’, ’hidden_layer_sizes’: (500,), ’learning_rate_init’: 0.001
Agr: ’activation’: ’logistic’, ’hidden_layer_sizes’: (100,), ’learning_rate_init’: 0.001
Neu: ’activation’: ’relu’, ’hidden_layer_sizes’: (50,), ’learning_rate_init’: 0.05
Opn: ’activation’: ’tanh’, ’hidden_layer_sizes’: (50,), ’learning_rate_init’: 0.001
Con: ’activation’: ’relu’, ’hidden_layer_sizes’: (500,), ’learning_rate_init’: 0.005

ALBERT (Layer 11) + MLP
Ext: ’activation’: ’logistic’, ’hidden_layer_sizes’: (500,), ’learning_rate_init’: 0.001
Agr: ’activation’: ’logistic’, ’hidden_layer_sizes’: (200,), ’learning_rate_init’: 0.001
Neu: ’activation’: ’tanh’, ’hidden_layer_sizes’: (200,), ’learning_rate_init’: 0.005
Opn: ’activation’: ’logistic’, ’hidden_layer_sizes’: (100,), ’learning_rate_init’: 0.001
Con: ’activation’: ’logistic’, ’hidden_layer_sizes’: (200,), ’learning_rate_init’: 0.001

ALBERT (Layer 12) + MLP
Ext: ’activation’: ’relu’, ’hidden_layer_sizes’: (200,), ’learning_rate_init’: 0.05
Agr: ’activation’: ’logistic’, ’hidden_layer_sizes’: (500,), ’learning_rate_init’: 0.01
Neu: ’activation’: ’tanh’, ’hidden_layer_sizes’: (500,), ’learning_rate_init’: 0.005
Opn: ’activation’: ’logistic’, ’hidden_layer_sizes’: (100,), ’learning_rate_init’: 0.001
Con: ’activation’: ’logistic’, ’hidden_layer_sizes’: (500,), ’learning_rate_init’: 0.01

DistilBert (Layer 2) + SVM
Ext: ’C’: 5, ’epsilon’: 0.5, ’kernel’: ’rbf’
Agr: ’C’: 1, ’epsilon’: 0.2, ’kernel’: ’rbf’
Neu: ’C’: 5, ’epsilon’: 0.5, ’kernel’: ’rbf’
Opn: ’C’: 1, ’epsilon’: 0.2, ’kernel’: ’rbf’
Con: ’C’: 0.1, ’epsilon’: 0.5, ’kernel’: ’rbf’

DistilBert (Layer 5) + SVM
Ext: ’C’: 5, ’epsilon’: 0.5, ’kernel’: ’rbf’
Agr: ’C’: 5, ’epsilon’: 0.2, ’kernel’: ’rbf’
Neu: ’C’: 5, ’epsilon’: 0.5, ’kernel’: ’rbf’
Opn: ’C’: 5, ’epsilon’: 0.2, ’kernel’: ’rbf’
Con: ’C’: 5, ’epsilon’: 0.2, ’kernel’: ’rbf’

DistilBert (Layer 6) + SVM
Ext: ’C’: 1, ’epsilon’: 0.5, ’kernel’: ’rbf’
Agr: ’C’: 1, ’epsilon’: 0.2, ’kernel’: ’rbf’
Neu: ’C’: 1, ’epsilon’: 0.5, ’kernel’: ’rbf’
Opn: ’C’: 1, ’epsilon’: 0.2, ’kernel’: ’rbf’
Con: ’C’: 1, ’epsilon’: 0.2, ’kernel’: ’rbf’

ALBERT (Layer 2) + SVM
Ext: 'C': 5, 'epsilon': 0.5, 'kernel': 'rbf'
Agr: 'C': 5, 'epsilon': 0.2, 'kernel': 'rbf'
Neu: 'C': 5, 'epsilon': 0.5, 'kernel': 'rbf'
Opn: 'C': 5, 'epsilon': 0.2, 'kernel': 'rbf'
Con: 'C': 5, 'epsilon': 0.2, 'kernel': 'rbf'

ALBERT (Layer 11) + SVM
Ext: 'C': 1, 'epsilon': 0.5, 'kernel': 'rbf'
Agr: 'C': 5, 'epsilon': 0.2, 'kernel': 'rbf'
Neu: 'C': 1, 'epsilon': 0.5, 'kernel': 'rbf'
Opn: 'C': 1, 'epsilon': 0.2, 'kernel': 'rbf'
Con: 'C': 0.1, 'epsilon': 0.5, 'kernel': 'rbf'

ALBERT (Layer 12) + SVM
Ext: 'C': 20, 'epsilon': 0.5, 'kernel': 'rbf'
Agr: 'C': 1, 'epsilon': 0.2, 'kernel': 'rbf'
Neu: 'C': 1, 'epsilon': 0.5, 'kernel': 'rbf'
Opn: 'C': 1, 'epsilon': 0.2, 'kernel': 'rbf'
Con: 'C': 0.1, 'epsilon': 0.5, 'kernel': 'sigmoid'

DistilBert (Layer 2) + GP
Ext: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)

DistilBert (Layer 5) + GP
Ext: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)

DistilBert (Layer 6) + GP
Ext: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)

ALBERT (Layer 2) + GP

Ext: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)

ALBERT (Layer 11) + GP
Ext: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)

ALBERT (Layer 12) + GP
Ext: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Agr: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Neu: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Opn: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)
Con: 'kernel': RBF(length_scale=1) + WhiteKernel(noise_level=1)

DistilBert (Layer 2) + AdaBoost
Ext: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 50
Agr: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 300
Neu: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 300
Con: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 100

DistilBert (Layer 5) + AdaBoost
Ext: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 50
Agr: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 300
Neu: 'learning_rate': 0.01, 'loss': 'square', 'n_estimators': 50
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 300
Con: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 150

DistilBert (Layer 6) + AdaBoost
Ext: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 200
Agr: 'learning_rate': 0.001, 'loss': 'square', 'n_estimators': 150
Neu: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 100
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 200
Con: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150

ALBERT (Layer 2) + AdaBoost
Ext: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 50

Agr: 'learning_rate': 0.01, 'loss': 'square', 'n_estimators': 200
Neu: 'learning_rate': 0.1, 'loss': 'linear', 'n_estimators': 50
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Con: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 100

ALBERT (Layer 11) + AdaBoost
Ext: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 100
Agr: 'learning_rate': 0.01, 'loss': 'square', 'n_estimators': 50
Neu: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 300
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Con: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 50

ALBERT (Layer 12) + AdaBoost
Ext: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 300
Agr: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 200
Neu: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 150
Opn: 'learning_rate': 0.01, 'loss': 'linear', 'n_estimators': 200
Con: 'learning_rate': 0.001, 'loss': 'linear', 'n_estimators': 300

Lars-Magnus Underhaug

From Traits to Threats

NTNU
Norwegian University of
Science and Technology