Master's thesis

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

Ingrid Vrålstad Løvås

Recognizing Social Media Right-Wing Radicalization

Using Text Analysis and Artificial Intelligence

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

NTNU
Kunnskap for en bedre verden

Ingrid Vrålstad Løvås

Recognizing Social Media Right-Wing Radicalization

Using Text Analysis and Artificial Intelligence

Master's thesis in Computer Science
Supervisor: Björn Gambäck
June 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

Right-wing extremists perform and plot most acts of extremism related violence in the Western world. Nevertheless, most research is done on religious extremism. The rise of social media has given multiple benefits, like making it easy to stay in touch with friends and family. However, it has also given an extra and easily accessible arena for radicalization. This master's thesis focuses on recognizing right-wing radicalization on social media using artificial intelligence. Social media posts from Gab and Twitter were collected to be analyzed and used for natural language processing and training of artificial neural networks to classify posts as neutral, radical or extreme. The extreme and radical datasets were collected from Gab, while the neutral dataset was publicly available.

The three datasets were analyzed based on the language features: post length, frequently used words, mentions, hashtags and URLs. It was found that politically related content characterized extreme and radical posts. Term frequency-inverse document frequency (TF-IDF) was used to calculate how extreme a post is based on IDF scores valuing the 500 most frequently used words in the extreme dataset. It was considered most important to minimize the erroneous of neutral posts as either radical or extreme. A fourth dataset was created to evaluate the trained models containing equally distributed data between the neutral, radical and extreme categories. The best artificial neural network gave a recall score of 86.6% for the neutral classifications.

As a supplement to the presented tasks, the Norwegian police were contacted to explore their current work to prevent radicalization. Artificial intelligence is not used in Norway for surveillance online to recognize people vulnerable to radicalization due to such as the importance of freedom of speech and privacy rights.

Sammendrag

I den vestlige verden blir de fleste voldelige ekstremismerelaterte hendelsene utført av høyreekstreme. Likevel er mesteparten av forskningen på ekstremisme relatert til religiøs ekstremisme. Økningen og utviklingen i bruk av sosiale medier har ført med seg mange fordeler, som at det har blitt enklere å holde kontakten med venner og familie. Dessverre har ikke sosiale medier bare ført med seg positive ting; det har også oppstått en ny arena for selv-radikalisering, rekruttering av nye medlemmer til ekstreme grupper og muligheten for å tilegne seg nye, ekstreme meninger. Denne masteroppgaven fokuserer på å gjenkjenne høyreradikal radikalisering på sosiale medier ved bruk av kunstig intelligens. Poster publisert på plattformene Gab og Twitter ble samlet og analysert. Videre ble dataen brukt til språkprosessering og for å trene kunstige nevrale nettverk slik at postene kunne bli klassifisert som nøytrale, radikale eller ekstreme. De ekstreme og radikale datasettene ble hentet fra Gab, mens det nøytrale datasettet inneholdt tweets og er offentlig tilgjengelige.

De tre datasettene ble analysert basert på postlengde, mest brukte ord, referanser til andre personer, hashtagger og linker. Resultatene viste at politisk innhold karakteriserte de radikale og ekstreme postene. TF-IDF ("Term Frequency-Inverse Document Frequency") ble brukt til å beregne hvor ekstrem en post er basert på IDF-verdier gitt ut i fra hvor mye brukt hvert ord var i det ekstreme datasettet. Den viktigste evaluering av maskinlæringsmodellene var hvor godt den klarte å minimere feilklassifisering av nøytrale poster som radikale eller ekstreme. Et fjerde datasett, som inneholder lik fordeling av nøytrale, radikale og ekstreme poster, ble laget for å evaluere de trente modellene. Det beste kunstige nevrale nettverket ga en recall score på 86.6% for den nøytrale klassen.

Det norske politiet ble kontaktet for å få innsikt i hva som gjøres for å forebygge radikalisering. Kunstig intelligens brukes ikke av det norske politiet til overvåkning av tilfeldige personer på nett eller for å oppdage personer som er sårbare for å bli påvirket av ekstremisme. Politiet driver ikke med overvåkning siden ytringsfriheten og retten til privatliv forhindrer dette.

Preface

The master's thesis is the final assignment needed to get a Master of Science in Computer Science from NTNU. The thesis is written in the 10th semester, the last one. As a preparation for the master thesis, a specialization project is written in the 9th semester. The specialization project serves as an exercise in writing an academic thesis and an opportunity to get familiar with the research field for the master's thesis. This thesis references the specialization project multiple times. The goal of the specialization project was to prepare for this master's thesis by getting familiar with work done to predict people vulnerable to social media extremism using artificial intelligence. Björn Gambäck supervised both the specialization project and the master's thesis.

Fellow student Thomas Alejandro Ramirez Fernandez contributed by gathering data and collaborating to get to know the field of social media extremism. He has contributed with valuable discussions and ideas.

Special thanks to the Norwegian police that agreed to conversations with Fernandez and the author, and for delivering information motivating the master's thesis.

Ingrid Vrålstad Løvås
Trondheim, June 9, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this section, the motivation and background for this master's thesis will be presented, followed by the goals and the research questions for the project. Then the research method used in the project is introduced. Some definitions specifically defined for the experiments and discussions in the thesis are then explained. The contributions of this work are presented before finally introducing the structure of the thesis.

1.1 Motivation

Radical beliefs are becoming more visible than before through the internet, especially on social media platforms. In addition, the internet and social media give an extra platform for self-radicalization and recruitment by extreme groups. During the Covid-19 pandemic, we have again seen how information is spread that suggests mistrust of institutions like the government. Mistrust in the government is a typical factor that can start a radicalization process [Angus, 2016]. Another example is the political polarization that has embossed the news, especially the time around the US election. Political polarization means that the population is divided into "poles" that have opposite beliefs, typically either far-right or far-left [Pew Research Center, 2014]. Radical beliefs are often expressed on social media platforms that easily is spread to other like-minded people.

The specialization project showed that most research on social media radicalization is concerned with religious extremism, especially Islamist extremism. Jones et al. [2020] stated that from 1944 to 2020, 57% of the performed terror attacks and plots were performed by right-wing terrorists, while only 15% were motivated by religious extremism. Since right-wing extremism is considered a significant threat in the Western world, it was decided that focusing on right-wing extremism would be rewarding and necessary. The specialization project additionally revealed that the most relevant literature focused on classifying users or posts as either extreme or not extreme rather than predicting future extremists or people vulnerable to entering the radicalization process. Therefore, it was decided to base the master's thesis on investigating the process of people having common beliefs, acquiring radical beliefs, and finally turning to right-wing extremists by examining their posts. The meaning of ordinary, radical and extreme users for this thesis will be explained in section 1.4.

During the work, several meetings were held with the Norwegian police. The police are restricted in their work by laws but can investigate tips from others. Suppose it was possible to develop a method for predicting if someone is in the radicalization process and likely to become an extremist. It could be used to tip the police but not by the police themselves. That suggests the importance of working with methods to predict right-wing extremists, and statements from the contact person in the police expressing the importance of such work boosted the motivation. It was expressed that the police could benefit from private individuals or organizations working to find effective and accurate methods to recognize vulnerable people so the police could investigate people further.

The definition of extremism presented in section 1.4 defined for this work states that an extremist does not have to pose a violent danger to himself, others or the society, but a danger of developing a politically polarized society. It is essential to state that for the police to act, there has to be a real danger for the reported person performing an act of violence. Suppose the detected vulnerable people do not pose a violent threat but may contribute to a polarized society. In that case, the police do not have the authority to act on the tip. However, the police can assist in implementing prevention measures in society. The limit of police permissions motivated the need to investigate alternative approaches to prevent radicalization.

1.2 Goals and Research Questions

In this section, two goals for the work are presented. The research questions (RQs) describe the tasks that should be performed to achieve the goals.

Goal 1: *Investigate differences between right-wing extremists' and other users' language used on social media based on the collected datasets.*

A dataset containing right-wing social media content and a dataset with neutral social media content are needed to achieve the goal. Additionally, it was desirable to collect a radical dataset. Two RQs were formulated, which, when answered, should help to fulfil the first goal.

RQ 1: *Can appropriate datasets for social media content for extreme and radical right-wing content and other users be found publicly available or collected from social media platforms?*

Data is needed to analyze the differences in language used by extremists and other users. When RQ 1 is answered, the needed material for answering RQ 2 exists.

RQ 2: *Do extreme right-wing users use language features different than other users?*

When RQ 2 is answered, the first goal should be fulfilled. The investigated language features were frequently used words, hashtags, mentions, post length and URLs. Thoroughly understanding the different language use is an enormous task, and therefore it is considered good enough to investigate the mentioned features.

Goal 2: *Use natural language processing and machine learning to recognize people in a radicalization process of turning into an right-wing extremist.*

To reach the goal, word embedding techniques should be implemented to see patterns in the language and similarities in the usage of terms. Machine learning should be used to train a model to find people vulnerable to entering the radicalization process or people approaching the point of turning into an extremist. The goal is to investigate right-wing extremism, not other types of extremism. RQ 3 and RQ 4 are researched separately, meaning that they do not use the other's results.

RQ 3: *Can natural language processing techniques be used to calculate a degree of how extreme a post is?*

Examining the frequency of used words in the extreme dataset using a variant of TF-IDF should be tried to collect a dictionary of important, typical right-wing words used by extremists. Further, it should be checked if the extracted words could calculate accurate results suggesting how radical a social media post is. Word2Vec should be used to examine the similarities between used words and if it could be used to capture the context.

RQ 4: *Can artificial neural networks be trained to classify if posts are neutral, radical or extreme?*

It is desirable to automatically recognize people in danger of becoming radicalized, which most likely is comprehensive when performed manually. Therefore a task of the master's thesis was to implement a machine learning model, and it was decided to try using artificial neural networks to classify posts into three different degrees of radical, namely neutral, radical and extreme.

1.3 Research Method

This section gives an overview of the approach performed during the project's work, which can be divided into two parts. The majority of the research was concerned with analyzing the collected data and applying natural language processing and machine learning to the data to create predictive models. Firstly, it was collected data from Gab and Twitter, followed by an analysis of the language features. After that, it was chosen and used natural language processing techniques. Before discussing the results and suggesting future work, a machine learning method was chosen, and the chosen model was trained. The approach for the described tasks is presented in chapter 5.

In addition to performing the described research, it was desired to get in contact with the Norwegian police. Possible conversations aimed to acquire insight into how the government works to prevent radicalization, detect extremism, and use AI as a tool for these tasks. Since the objective of the master's thesis is highly relevant for governmental instances, it was figured that it would be worth while emailing the Norwegian police[1], the police security service (PST)[2], the police ICT services[3] and the police directorate[4]. Unfortunately, none of the emails resulted in further communication. In the last few years, the Norwegian police have become more present on social media, like Instagram. The Oslo police district was therefore contacted on Instagram and responded immediately. The Instagram communication resulted in contact with a helpful woman in the ICT and business development section of the Oslo police district. In the following chapters, she will be referenced as our contact person. Three meetings were arranged with different units of the police. None of the participants in the meetings will be mentioned by name in this thesis. The communication is described in section 2.4.

[1] The Norwegian police's webpage: `https://www.politiet.no`

[2] PST's webpage: `https://www.pst.no`

[3] The police ICT services contact information: `https://www.politiet.no/om/organisasjonen/sarorganene/pit/`

[4] Contact information of the police directorate: `https://www.politiet.no/om/organisasjonen/sarorganene/politidirektoratet/`

1.4 Specific Definitions

To justify the project's results, it is necessary to define the terms specific to this task. For the following terms, it exists formal definitions, but these are all open for interpretation and specification. A few examples of existing definitions are presented in chapter 2. It is essential to define these terms since it may give significant misleading results if the understanding of the terms is not correct. The following definitions apply only for the master's project, not for the related work in section 4.2.

1.4.1 Extremism

In subsection 2.1.1 two formal definitions will be presented that include different types of extremism as political, ideological and religious extremism. Since this project is concerned with right-wing extremism, it was decided to refer to extremism as right-wing extremism for the rest of the thesis. Then it is specified that the results in the project may not be adaptable to other types of extremism. In many situations extremists are considered people who may perform acts of violence, but in this thesis, extremism is considered a broader range of people. A hypothesis is that people with right-wing beliefs pose a danger to society by expressing their beliefs without performing any physical actions. The danger of getting a more politically polarized society can, in the long term, be even more dangerous than violent acts because of the number of people that stands together and how long the people have been exposed to the extreme content. It is considered a concern that the world, or parts of the world, may reach a point where it is too late to prevent a polarized society resulting in a society that cannot communicate and work together as valued in a democracy.

Wolfowicz et al. [2021] presents that fewer than 1% of extremists turn into terrorists, meaning that the majority of people with extreme beliefs will not perform actions of violence. Therefore, it is necessary to investigate people with extreme beliefs in general, contrary to the Norwegian police that states that the extremist needs to be a violent threat for these experiments. The definition of extremists for this task is people who express beliefs that are not considered normal. Especially beliefs concerned with race, religion, sexual orientation, gender roles or mistrust in the government or other institutions. In the experiment setup in subsection 5.2.2, it is more specifically defined the criteria for being considered extreme in this project.

1.4.2 Radical

This project decided to define radical as beliefs that are not normal but not extreme enough to classify a person or social media post as extreme. Subsection 5.2.2 explains how it was decided what is radical. The definition of radical is based on the collected radical dataset.

1.4.3 Other Users

In this thesis, a user is referred to as either extreme, radical, or an other user. The other users are sometimes referred to as neutral users. When stating that a user belongs to the latter group, it means that it does not belong to the extreme or radical dataset and does not fulfil the criteria specified to be defined as extreme or radical. It can not be guaranteed that the other users do not hold radical beliefs.

1.4.4 Activation Point

The activation point in this project is considered the event when someone goes from only being in the radicalization process to becoming an extremist. It is important to recognize the people considered vulnerable to radicalization before they reach the event of becoming an extremist. It is hard to define when someone reaches this point. How can we decide and define when someone is an extremist? Who has the power to decide this? In this project it was decided to divide social media post into three categories, namely *Extreme*, *radical* and *neutral* to help defining the activation point.

1.5 Contributions

The listed points are the contributions the research of this master's thesis delivers to the field.

1. *Gathering of two novel datasets from the social media platform Gab.*

2. *An analysis of the language used in three datasets classified, by the author, as right-wing extreme, radical and neutral.*

3. *A dictionary of right-wing words, with a suggested belonging value describing its importance in the collected extreme dataset.*

4. *A classifier for classifying social media posts into extreme, radical or neutral.*

1.6 Thesis Structure

The structure of this thesis is described below. The thesis consists of eight chapters, and the first one is the introduction chapter.

> **Chapter 2** gives an introduction to social media radicalization by first elaborating on extremism, radicalization, terrorism and freedom of speech, followed by an explanation of different social media platforms. Lastly, the chapter presents the information delivered in the meetings with units of the Norwegian police.
>
> **Chapter 3** delivers background theory about the technical techniques and methods used in the experiments. Some sub-fields of AI are presented, evaluation metrics are explained, and finally, the chapter presents technical tools used in the development process.
>
> **Chapter 4** presents the method used to collect relevant literature in the specialization project and the findings of the literature review.
>
> **Chapter 5** presents the plans for the experiments and the setup for the data gathering, text analysis, preprocessing of data and the artificial neural network multi-label classifier.
>
> **Chapter 6** explains and visualizes the findings of the experiments explained in chapter 5.
>
> **Chapter 7** discusses the goals and RQs based on the information given by the police, the relevant literature and the results from the experiments. The chapter includes a section discussing the findings not covered by the research goals and RQs. After that, the approaches and the results of the project are evaluated.
>
> **Chapter 8** lastly concludes the master's thesis and suggests future work that should be explored.

Chapter 2

Domain Background

The chapter presents the domain theory needed to understand the need for the thesis and the experiments performed. Domain means information about such as social media, extremism, radicalization and freedom of speech. In addition, the chapter presents information about the units of the police that were present in the meetings (see section 1.3) and the information retrieved during those meetings. Section 2.3 is taken from the specialization project.

2.1 Definitions

The four terms *Radicalization*, *Extremism*, *Terrorism* and *Freedom of Speech* are all relevant to discuss when investigating and implementing preventive measures, but they have different definitions. Definitions presented by others are presented in this section, while the specific, self-defined definitions for this specific project were presented in section 1.4. Defining extremism can be challenging since the variations of extremism can vary in degree of danger. Some could even argue that it does not have to be dangerous for anyone. Beliebers[1] can be perceived as extremes because of their devotion to an unknown person. Despite this, a belieber is generally not dangerous to anyone. Suicide, eating disorders, and drug addiction can be classified as types of extremism. In comparison, others would primary mention religious extremism if asked. In the specialization project, religious and political extremism were introduced before it was decided to focus on political extremism, specifically, right-wing extremism.

[1] Beliebers is Justin Bieber's fans. `https://no.wikipedia.org/wiki/Beliebers` (Accessed: 22.04.22)

2.1.1 Extremism

One existing definition of extremism is "political, religious, etc. ideas or actions that are extreme and not normal, reasonable or acceptable by most people", which is presented by Oxford Learner's Dictionary [2021]. A second definition is "the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty and the mutual respect and tolerance of different faiths and beliefs. We also regard calls for the death of members of our armed forces as extremist." which is presented by the HM Government [2015, p.9] of the United Kingdom in their report *Counter-Extremism Strategy.* The Norwegian police elaborated on their definition of extremism which states that for the police to act, the individual has to pose a violent danger. The content given by the various conversations with units of the police is further elaborated in section 2.4. They expressed the importance of explicitly defining used terms, especially extremism, before starting to work.

2.1.2 Radicalization

Radicalization is a process that can be sequential or not, from when someone has common beliefs to starting to acquire radical beliefs to becoming an extremist. The process does not necessarily lead to the person becoming an extremist [Regjeringen.no, 2019]. This project focused on the whole process, from not having radical beliefs to entering the radicalization process till they pass the activation point turning into extremists. It is desirable to recognize the users prior to the activation point since it is easier to implement preventing measures than to deradicalize already extremists.

2.1.3 Terrorism

Radicalization and extremism can lead to terrorism. Terrorism is defined as an act that is meant to influence or scare the government, international institution or the public [Legislation.gov.uk, 2021], and as radicalization, it can happen in ideological, political and religious cases. Acts of terror often lead to deaths or wounded people, while extremists do not have to perform actions that directly affect others. The focus of this thesis will not primarily be on terrorism but more on acquiring radical beliefs that lead to a polarised society, which of course, may lead to acts of terror. However, a polarized society can be at least as dangerous but not so visible to ordinary people. Polarization is defined as the distance in ideology between two opposing political parties [Milaĉić, 2021]. Milaĉić [2021] states that a degree of polarization is a necessity for democracy since it increases the interest in politics in the population, but if the distance grows too much, it poses a threat.

2.1.4 Freedom of Speech

The creation of *Universal Declaration of Human Rights* is considered a milestone in history for human rights. It was acknowledged by the United Nations on the 10. of December in 1948. Article 19 of the declaration is concerned with the right to freedom of speech [UN General Assembly, 1948].

> **Article 19:**
> *Everyone has the right to freedom of opinion and expression; this right includes the freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.*

As The Norwegian Government [2020] states, freedom of speech is a prerequisite for democracy. The next question is what is included in *freedom of speech*? Does it mean that you are allowed to express exactly what you mean without caring about other people? Is it allowed to express hateful beliefs or discriminatory and racist statements?

> *Freedom of expression is not only a prerequisite for democracy, it is also vital for the realization of other fundamental human rights, such as freedom of assembly and freedom of religion or belief.*
> - The Norwegian Government [2020]

The definition of freedom of speech for Amnesty International UK [2020] gives no limits to what is legal to express. Different countries have different limitations on the right to freedom of speech.

> *Freedom of speech is the right to seek, receive and impart information and ideas of all kinds, by any means.*
> - Amnesty International UK [2020]

The countries that are committed to the UN declaration of Human Rights need to have statutory provisions to make exceptions from article 19.

2.2 Right-Wing Extremism

In the literature review presented in section 4.2, it was discovered that most previous research was done on religious extremism even though right-wing terrorists perform most acts of terror in the western world. From 1994 to 2020, in the United States, 57% of the acts of terror and plots were performed by right-wing terrorists, while religious terrorists perpetrated only 15%. The statistics were delivered by the Center for Strategic and International Studies (CSIC) [Jones et al., 2020, p.2]. Even the number of acts (25%) performed by left-wing extremists exceeds the religious acts. The numbers only represent how many cases each group performed, but because of the 9/11 attack, the most fatalities were because of religious extremism from 1994 to 2020. However, it substantiates the claim that to investigate how to prevent political, predominantly right-wing, radicalization is needed.

Since the focus is right-wing extremism, subsection 2.2.1, subsection 2.2.2 and subsection 2.2.3 are included to give insight into the "typical" members, victims and social media language.

2.2.1 Members

A typical member of right-wing communities is a white, middle-aged man that values traditional values like the stereotypical gender roles. The Norwegian police security service (PST) presented in their report from 2019 about the backgrounds of people in extreme right-wing communities that men are over-represented in these communities due to the idealization of the masculine [PST, 2019, p.7]. Members can be such as neo-Nazis, white supremacists, incels, anti-government, anti-immigration and racists. The people that support Hitler's nazi politics and ideas are called neo-Nazis [Britannica Dictionary, 2022]. The white supremacists think that the white race is superior to others, and incels are defined as men that are involuntary celibates and possibly perform acts of violence against women [Jones et al., 2020, p.5-6].

2.2.2 Victims

Typical victims of right-wing extremism are non-white people, Jews and Muslims [Ravndal et al., 2020]. The right-wing expression also targets people in the LGBTQ+ communities. The type of right-wing extremism varies in different geographic locations and different cultures. Hence it can be challenging to recognize international patterns of victims and expression manner because of the major variations around the world.

2.2.3 Language

Right-Wing extremism uses different numbers, slang etc., to express their opinion in addition to using ordinary language. Fraštíková and Demčišák [2019] found in their research that right-wing text contained a higher degree of refugee and immigration content than for other users. Their research showed that right-wing people also tended to write content concerned with sceptical views of the EU, patriotism, polarization, and anti-elitism. Right-wing users often use humor and

	Symbol	Explanation	Source
1	*((( )))*	People mentioned by name inside three parenthesis are considered Jewish.	[2017]
2	*cuck*	If someone is refereed to as a cuck they are considered to be brainwashed or ignorant.	[2017]
3	*Culture enricher*	Is used instead of immigration or an immigrant.	[2021]
4	*88*	Stands for two times H, the eighth letter in the alphabet, representing the two words "Heil Hitler".	[2018]
5	*18*	The first and eight letters are AH, which by the right-wing symbolize "Adolf Hitler".	[2018]
6	*14*	It represents the 14 words long slogan: "We must secure the existence of our people and a future for white children".	[2018]
7	*RaHoWa*	Stands for "Racial Holy War" which is used by white supremacist symbolizes the confrontation with the non-white world.	[2012]

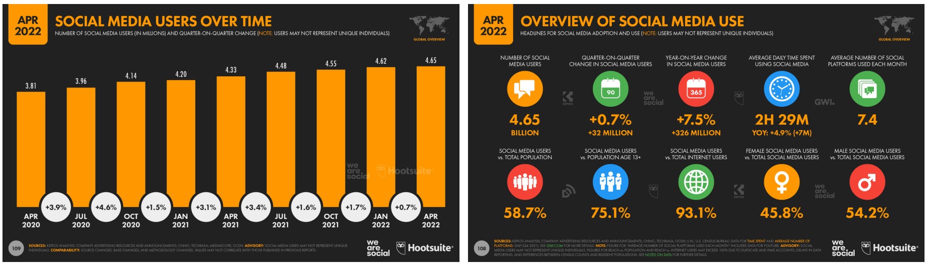
Table 2.1: Examples of terms used by right-wing.

sarcasm to cover their beliefs to appear less extreme [Hervik, 2019]. Dog whistling is commonly used by the right-wing, meaning that they cover their beliefs using other words resulting in only fellow right-wing people understanding the message [Haney-López, 2014]. Memes are also used to cover up a message, making it harder to recognize extreme right-wing content [Daniels, 2018]. Some typical symbols and words of the right-wing's social media language are displayed in Table 2.1. The 1 and 2 entry in Table 2.1 are presented by Marwick and Lewis [2017], number 3 by Åkerlund [2021], 4-6 by Bundesamtür Verfassungsschutz [2018] and the last by IACP [2012].

2.3 Social Media

This section is taken from the specialization project with only minor language, content changes and updated numbers of the digital statistics.

Investopedia defines *social media* as "... a computer-based technology that facilitates the sharing of ideas, thoughts, and information through virtual networks and communities." [Dollarhide, 2021]. Social media includes platforms where text, pictures, videos, and links are shared. The specialization project and the



(a) Growth social media users 04.22-04.22.

(b) Overview of social media use.

Figure 2.1: Social media status April 2022.
[Kemp, 2022]

master focuses on platforms containing short and informal text. Therefore, this section presents Facebook and Twitter. In addition, an insight into the use of forums and blogs are included. Lastly, presenting other relevant platforms that claims they value free speech, typically attracting users holding radical or extreme beliefs.

The *Digital 2022 April Global Statshot Report* presents that the number of social media users has passed 4.6 billion users, as shown in Figure 2.1a [Kemp, 2022]. Following the current trend, 60% of the world's population will be on social media by the first half of 2022 was predicted by Kemp [2020] in the *Digital 2021 October Global Statshot Report* and by the first quarter of 2022 the percentage was 58.7% according to Kemp [2022]. Kemp [2022] presents that the average user uses 6 hours and 53 minutes on the internet, where 2 hours and 29 minutes are on social media as Figure 2.1b shows. Hence, spending a lot of our waking time in front of a monitor exposed to good and bad content.

2.3.1 Facebook

Facebook[2] is a website and mobile application that delivers an online society. It allows its users to connect with friends, chat with them and share content. In addition, it supports the creation of public and private groups. Facebook is a free platform financed by ads. Therefore the user is constantly fed with involuntarily content. The user can post status updates with friends containing text, pictures, URLs, and other less important and used features. Facebook allows its users to write long texts containing up to 63 206 characters in each post [Zote, 2021]. Kemp [2020] predicted that Facebook would be the most used

[2] The Facebook website: `https://www.facebook.com`

social media platform in 2021. Meta Platforms, Inc [2022] released a report for the first quarter of 2022 stating that Facebook, in April 2022, have 1.96 billion daily active users which was an increase from 1.93 in September 2021 [Facebook, 2021].

2.3.2 Twitter

Twitter[3] is a microblogging platform that allows users to create a personal profile, post short status updates called *tweets*, follow other profiles, read their tweets and re-post others' tweets. Twitter has a limit of 280 characters forcing its users to share concise updates [Twitter, 2021]. As of July 2021, Twitter was the 16th most popular social media platform in the world based on the number of users[4]. In the United States, Twitter was the fourth used platform as of 2019[5]. In Twitter's earnings report to shareholders for the third quarter in 2021, they reported that Twitter had 211 millions average monetizable daily users [@TwitterIR, 2021, p.2]. The first quarter report of 2022 from Twitter presented that the average monetizable daily users had increased to 229 millions [Twitter, Inc, 2022].

2.3.3 Forums and Blogs

Forums are platforms that allow for discussions, where every user is considered a contributor. They can be forums like Twitter that do not target a specific audience or subject, or they can concentrate on a specific subject. It is convenient for different communities, in extreme cases, for example, extreme communities. One person or group/organization usually runs a blog. The owner typically uses the blog to front their case and opinions. Usually, the blog allows its readers to comment on a post but not post their content.

2.3.4 Free speech platforms

Both Twitter and Facebook have started to remove fake news and extreme content, and in some cases, ban users[6]. Therefore, many radicals have moved to other platforms where users rarely get banned even if radical beliefs are expressed.

[3] The Twitter website: `https://twitter.com/`

[4] Most popular social networks worldwide as of July 2021, ranked by the number of active users: `https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/` (Accessed: 03.11.21)

[5] Most popular mobile social networking apps in the United States as of September 2019, by monthly users: `https://www.statista.com/statistics/248074/most-popular-us-social-networking-apps-ranked-by-audience/` (Accessed: 03.11.21)

[6] The Twitter rules are available at `https://help.twitter.com/en/rules-and-policies/twitter-rules` (Accessed: 08.12.21). Facebooks community standards are presented at `https://transparency.fb.com/policies/community-standards/` (Accessed: 08.12.21).

Some examples of more radical social media platforms are Gab[7], Vkontack[8] and Stormfront[9].

2.3.5 Gab

Gab is similar to Twitter but expresses how important freedom of speech is, as their LinkedIn profile states. Its profile describes the platform with these words: "Our mission is to defend, protect, and preserve free speech online for all people."[10]. In comparison to Twitter, the length of the gabs is long. In the start, the character limit was 300, similar to tweets, but a gab can now consist of 3,000 characters[11]. Due to their extreme expressions, many Gab users have been banned from the most popular social media, typically Twitter. Gab allows QAnon conspiracy, anti-Semitism, misinformation, general hatred and racist content.

2.3.6 VKontact

VKontact (VK) is a Russian social media platform similar to Facebook. It is the most used social media platform in Russia, and it had 97 million active monthly users in November 2021[12]. As Gab, Vkontact has a higher threshold for banning users and removing content than Twitter and Facebook. Therefore it has become a popular platform for white supremacists.

2.3.7 Stormfront

Stormfront is the oldest online forum for white supremacists [ADL, 2021]. It is also considered one of the largest online communities for white supremacists. The founder of Stormfront, Don Black, has written in the post regarding the guidelines for posting that "Our mission is to provide information not available in the controlled news media and to build a community of White activists working for the survival of our people."[13].

[7] The Gab website: `https://gab.com/`

[8] The VKontact website: `https://vk.com`

[9] Stormfront's website: `https://www.stormfront.org/forum/`.

[10] Gab's LinkedIn profile can be found here: `https://www.linkedin.com/company/gab-ai-inc./about/` (Accessed: 03.11.21)

[11] Gab's post guide: `https://help.gab.com/article/basics-post-composer-options` (accessed: 01.05.22)

[12] VKontact's explanation of themselves: `https://vk.com/about` (Accessed: 03.11.21)

[13] The post *Guidelines for Posting* on Stormfront.org can be found here: `https://www.stormfront.org/forum/t4359/` (Accessed: 08.12.21).

2.4 Communication with the Norwegian Police

During the semester, Fernandez and the author held three meetings with different units of the Norwegian police. In the last one, a representative from the Dutch police attended. Firstly the section presents the meeting structure. Then a subsection for each meeting delivers an introduction to the attending unit and presents the relevant retrieved information. This section is included in the chapter about domain background since the police delivered insight into information considered to supplement the presented theory in this chapter. It is essential to state that this section consists of work done in this thesis, not just theory as in the previous parts of this chapter.

2.4.1 About the Meetings

Our contact person organized three meetings with people she tough would be relevant for us to communicate with. All meetings were held on Teams, and minutes from the meetings were written but is not shared unless asking the author for permission. None of the sessions was recorded. All the given information is publicly available. This section presents the meetings with the "Hatkrimgruppen", Kripos and a meeting of representatives with different occupations in the police and a representative from the Dutch police. The Dutch police was represented due to the Netherlands' position when it comes to implementing AI in the police.

2.4.2 "Hatkrimgruppen"

The first meeting our contact person organized was with the group of hate crimes located in Oslo. The section presents the reason for creating the group, general information about hate crime, and a conversation summary. Our contact person and a representative from the group of hate crimes attended the meeting.

About the Group of Hate Crimes

The group of hate crimes was established in 2014 and is located in the Oslo police district. All employees in the hate crime group have an education from the police, but none have an IT education. Instead, they have close collaboration with other units in the police that have specific knowledge, like IT. Oslo is the only police district in Norway with a unit exclusively working with hate crimes. In October 2021, a national competence center for hate crimes was created. The center is located in Oslo because of the Oslo police districts size, but it is a national unit. The center's tasks are to build competence in the police and assist all police districts in Norway with both knowledge and guidance in hate crime-related cases. Below is the translated definition of hate crime made by the Norwegian police [Oslo politidistrikt, 2022].

> Hate crimes are criminal acts that are wholly or partly based on the skin color of others, national or ethnic origin, religion/belief, sexual orientation, gender expression and gender identity, and/or disability.

Anti-Semitism, anti-Gypsyism, and hatred against the Sami people are included in the definition of ethnic origin. Additionally, the definition of hate crime includes national and ethical origin and skin color [Oslo politidistrikt, 2022].

The representative from the group of hate crimes said that for a statement to be classified as hate speech by the police, it has to be published/happen in a public place. The action/statement needs to reach at least 20 people, such as a Facebook group, or be observed by at least one person, like in a group chat in Messenger. Nevertheless, actions or statements that do not meet those criteria can be classified as hateful.

Motivation

The report by Oslo politidistrikt [2022] states that the number of classified hate crime reported cases in 2021 in Oslo was 325, which is an increase of 11% from 2020. Despite the increase, the number of reported cases on social media in Oslo decreased from 32 to 24 in the same period. Oslo politidistrikt [2022] was surprised that the number of cases decreased when the time spent by the average social media user increased during Covid. Oslo politidistrikt [2019] presented that the number of reported hate crime cases had increased by 94% from 2015, the first whole year of the group, till 2019, when the number of cases was 278. It is hard to say if the rise in reported occasions is due to more hate crimes or less dark numbers. The participant from the hate crime group said that they divide dark numbers into cases that are not reported to the police and reported cases that are not correctly identified as hate crime. The division of dark numbers for hate crimes is described by Politiet [2021] in their report about reported hate crimes in 2020. The rise in reports of occasions of hate crime, combined with the need to minimize the dark numbers, are motivation to continue the hate crime group's work and further development of the unit.

Report by Amnesty in Norway

A report presented by the Norwegian branch of Amnesty International about 21 political measures proposed to combat incitement online was discussed. Nine of the measures targeted the task of strengthening the preventive work of the police [Amnesty International Norway, 2019]. The nine measures relevant to the police were numbers 4-12, which are reproduced in English and presented in the list in this below.

4. Establish more hate crime groups in the police and hire more people.
5. Create a written mandate to combat illegal hatred online.
6. Strengthen the competence of the police.
7. Strengthen the attorney general competence and resources to work against illegal hate online.
8. Use the entire legislation actively.
9. Try using co-responsibility for legally use for criminal statements.
10. Focus on preventative work in the police.
11. Create more, and more detailed police statistics.
12. Ask the attorney general to evaluate the quality of the work by the police.

The police work with all the above measures and tasks defined by the police. During the meeting, the main focus was on measures four and six, which creates the basis for the rest of the points. The group of hate crimes have implemented the two measures. The personnel have increased as desired in point four, and the establishment of the national competence center was partly motivated by number six. Implementing co-responsibility, measure nine, for hate crimes is work in progress. Additionally, the police are working to acquire legal competence.

Approach

Prevention is a focus in the unit, but different parts of the police work to prevent radicalization and extremism. The hate crime unit does not use surveillance in their work, and surveillance is generally not performed by the police due to legal regulations. Individuals should not be perceived as suspicious before there is a significant reason to believe that the individual can pose a danger. The representative from the group of hate crimes stressed the danger of pre-censoring if performing surveillance and preventative measures. Freedom of speech is important in Norway, so the police should not impede the human right to express personal beliefs. Other institutions focus on preventative work. In subsection 2.4.3, it will be explained that the work of the police is to prevent acts of violence, not punish someone holding extreme beliefs. To distinguish between freedom of speech and illegal, hateful statements is hard, and therefore, the group of hate crimes works to get cases to be tried in the court so the verdicts can be used in future cases.

In the Penal Code ("Straffeloven"), the applied law for hate crime, the most used paragraphs by the group of hate crimes are §77 (subsection A.1) and §185 (subsection A.2). In 2021 it was added a new paragraph, §267a[14], to the criminal law that deals with sharing of videos of people in a vulnerable situation. The paragraph is starting to be used by the group, especially on videos shared of violence on social media. Especially if a case is taken to court and the trial's outcome is positive, it is desired to get publicity in the media to inform the Norwegian population about what is happening and educate the populating to easier know where the line between legal and illegal, hateful actions and statements goes.

2.4.3 Kripos

The second meeting was held with two representatives from Kripos, responsible for tracking extremism and radicalization. They are the two employees that work with extremism on social media in Kripos. Fellow student Fernandez also attended the meeting. This section presents some background information about Kripos before introducing their legal permissions and investigation approach.

Background

Kripos is a unit of the police that is organized under the Norwegian police directorate. The unit is national and has the responsibility of combating organized and other severe criminality [Politiet Kripos, 2022]. Kripos has existed since 1959[15].

One representative was a police officer, and the other was employed as a civil political scientist. They expressed the advantage of having different competence, so they complemented each other due to their difference in points of view on an issue. They work together, but one focuses mainly on right-wing extremism and the other on Islamist extremism.

Permissions

During the meeting, it was expressed the necessity of having a clear definition of what extremism and extremists are for the exact problem since there exist multi-

[14] The paragraph has not yet been translated to English. The Norwegian version can be found here: https://lovdata.no/lov/2005-05-20-28/§267a

[15] Kripos' webpage: `https://www.politiet.no/om/organisasjonen/sarorganene/kripos/`

ple. An example is the difference between Kripos' definition and the definition for this thesis. Kripos defines extremists as people who pose a threat of performing an act of violence. In contrast, extremists do not need to perform a violent act for this thesis, but acquiring extreme beliefs can be enough. The definition of extremism for this thesis is explained in section 1.4. Kripos, and the rest of the police, have limited permissions due to legal provisions. As the hate crime group expressed, Kripos is afraid of performing pre-censoring since it is not the police's responsibility.

The police is an instance that shall prevent acts of violence, and therefore the definition of extremism is defined to deal with possible violent people. People who hold extreme beliefs but do not pose a danger of violence are defined as radical. According to Norwegian laws, the people being classified as radical by the police are not doing anything illegal. However, situations not considered violent can be considered illegal, such as recruiting new members to extreme groups. An example is that a member of a militant Islamist Sunni Muslim group in Norway was convicted in 2018 for terrorist recruitment to IS [Garvik and Stenersen, 2018].

Kripos can not work to prevent a polarized society directly, but they can assist in preventative work in society. The Norwegian Ministry of Justice and Public Security presented an action plan for preventing radicalization and violent extremism consisting of 30 measures [Justis- og beredskapsdepartementet, 2020]. One way the police can prevent a polarized society is to contribute to the implementation process of these measures. Measures seven and eight are cornered with the implementation of resources in school to prevent radicalization, which are preventative measures presented and discussed in subsection 7.3.4. PST has fewer restrictions regulated by the authorities than the police. It means that PST can perform some degree of surveillance of people in the radicalization process even though it is not suspected that they pose a violent threat. PST's permissions are regulated in the police law in paragraph 17b, which includes that the PST holds the responsibility to prevent and investigate violations of the hate crime paragraph, §185, in the Penal Code (see subsection A.2)[16]. Additionally, they hold the responsibility for preventing and investigating other violations regulated by the Penal Code[17].

Approach

Kripos do not perform surveillance but receive tips about people others fear are extremists or in the radicalization process. They operate and develop a platform

[16] §17b in the police law was not found translated to English, but the Norwegian version can be found here: https://lovdata.no/lov/1995-08-04-53/§17b (Accessed: 06.06.22).

[17] See footnote 16 for more responsibilities of the PST.

where the population can send in tips, either with a name or choose to be anonymous. However, the tipper is taken into account when investigating. If the tipper is someone close to the suspect, the tip is expected to be real. Local police districts can also request an evaluation of a person, group or platform. Sometimes people in the same communities as known extremists are investigated, but it is rarely done. They use open-source online platforms to investigate the suspect when receiving a tip, and Facebook is the most used platform. The investigating process is complicated due to the blurry difference between illegal and legal behavior and between a radical person that will perform an act of violence and one who will not. It is tried to check if a person is vulnerable to radical influence or if they already can be defined as extremists. Indicators defined by domain experts are used in that process.

Suppose Kripos decides that a person poses a danger, they send a message of concern to the person's local police district, which chooses if they want to act on it. In Norway, measures can not be forced on a person, so it needs to be in voluntary cooperation with the suspect to implement measures to prevent a person in the radicalization process from becoming an extremist. Only if the police fear and believe that the person poses an immediate danger the police is allowed to act without the consent of the person because it can be classified as an act in self-defence cf. §18 (subsection A.3) in the Penal Code.

Kripos needs to be convinced that the person poses a danger to himself, others, or the society to not accuse people of not being in the radicalization process of being extreme. To the knowledge of the two representatives, it does not exist implemented software in the police to ease the detection process. Kripos is allowed to investigate peoples' public online platforms, but scraping data would be classified as a variant of surveillance. As desired for the experiments described in section 3.5, Kripos also want to reduce the false positive classifications. Since Norway allows for extreme beliefs and the expression of these beliefs in public, it is necessary to implement measures and beliefs in the society to prevent people from entering the radicalization process.

As mentioned, Kripos works with both right-wing and Islamist extremism, and the investigation approach is similar for both groups. Extreme Islamist content tends to be removed from social media faster than right-wing content. A reason could be that extreme Islamist content differs more from "normal" content than right-wing content.

2.4.4 Dutch Police

Our contact person organized the meeting to give the Norwegian police and us master students an insight into the status of the usage of AI in the Dutch police. Present in the meeting were a representative from the Police Lab AI in the Netherlands, two researchers working for the Norwegian police, the representative from the hate crime group, our contact person, Fernandez and the undersigned.

About

The Dutch Police Lab AI is a collaboration between the Dutch police, Utrecht University, University of Amsterdam and Delft University of Technology. On the webpage of the Innovation Center for Artificial Intelligence (ICAI), the Police Lab AI is described with the following sentence; *"They aim to develop state-of-the-art AI techniques to improve the safety in the Netherlands in a socially, legally and ethically responsible way."*[18]. The Netherlands is considered the leading police in Europe when it comes to the use of AI. The AI lab works to translate AI theory into practice. The representative from the Dutch police has a technical background. About 150 employees in the Dutch police are data scientists, showing the focus on the implementation of AI in police-related work.

Implementation of AI

The Dutch police have already implemented AI that is actively used. The first AI agent has assisted in over 300K cases. The agent delivers an assessment about if a case reported would be dismissed or accepted if submitted to the police. An example of a possible implementation of AI was in airports. If the police have reasons to confiscate a data carrier, such as a phone, then there is a limited window of time to sift through its data to look for indicators of radicalization. It could be a suitable place to implement fast methods to analyze the data to check for signs of radicalization. AI could be used to find the most important indicators of someone being in the radicalization process as a systematic approach, contrary to the current approach where intuition and psychological aspects are more used. It was also discussed that it could be promising to look at textual features and include audio and visual data.

Limitations

The Dutch police have similar restrictions as the Norwegian police, and the legal regulations limit how and where AI can be implemented. Surveillance of people is illegal for the Dutch police as well. To be allowed to gather data, the police

[18] ICAI's webpage for the Police Lab AI: `https://icai.ai/police-lab-ai/`

have to build a case and get an allowance to collect the data. Hence the Dutch police, like the Norwegian, could not implement the methods used in this project, but they could benefit from the results if people using such methods tipped the police about possible dangerous people.

2.4.5 Recommended Resources

During the period the meetings with the police were held, PST released a report about the national threat assessment for 2022 that our contact person recommended. The report states that Norway faces a moderate level of the terror threat and that right-wing extremists and extreme Islamist pose the highest threat. PST consider it likely that both groups will attempt to perform an act of terrorism in 2022. Right-wing terrorist threats can change quickly, according to PST. The radicalization of right-wing extremism will primarily be on social media platforms where people can exchange and share extreme content and contact other like-minded. PST explains that the content posted by right-wing users often dehumanizes minority groups like the ones described in subsection 2.2.2. The report also states that it is expected that extremists try to radicalize people in own circles, suggesting an advantage of using other data than just text, like followers, to predict vulnerable users [PST, 2022].

The police published a report about their threat assessment for 2022 where they state that it is expected threats and hateful statements will increase and that it can contribute to legitimizing these beliefs and statements and can have a radicalization effect on individuals [Politiet, 2022]. Politiet [2022] expect that the number of acts of violence will increase among minors, especially young boys and that social media can be a reason for the increase.

Chapter 3

Technical Theory

This chapter completes the needed background knowledge by supplementing chapter 2 by introducing technical theory. The technical theory includes explanations of artificial intelligence (AI), natural language processing (NLP) and machine learning (ML). There are introduced multiple NLP techniques and ML approaches. Additionally, technical tools used in the implementation are presented. Section 3.1, the paragraphs about n-grams and TF-IDF in subsection 3.2.1, section 3.3 and section 3.5 are taken from or based on the specialization project. Most of section 3.4 is taken from the specialization project, except subsection 3.4.3, where only the first paragraph is copied. Minor language or content changes may have been made.

3.1 Artificial Intelligence (AI)

There does not exist a consensus definition of AI. Multiple definitions exist, where the words *processing*, *reasoning* and *behavior* are used to define AI. AI is a subfield of computer science that aims to replicate human intelligence in smart machines [BuiltIn, 2021]. Some definitions define AI as a field trying to imitate human performance, while others want to strive for an ideal performance called rationality [Russel and Norvig, 2016, p.1]. Essential methods for NLP and ML will be presented in section 3.2 and section 3.4.

3.2 Natural Language Processing (NLP)

NLP is a part of AI that deals with making computers understand the text written and spoken by humans. Computational linguistics, rule-based models, statistic, and Machine Learning are combined to both interpret and understand human language [IBM, 2020]. Supervised learning was, in 2018, the most promising ML approach for NLP [Bengfort et al., 2018, p.8]. Natural language is how we write and talk. Contrary to formal language, there is no definite set that can be defined for sentences. Humans use different sentence structures, and words or sentences can have multiple meanings. In addition, the natural language is in constant development. In NLP a text is referred to as a *document*, while a collection of documents is called a *corpus* [Bengfort et al., 2018, p.19]. Since this project is concerned with short, informal texts, these documents are tweets and gabs.

3.2.1 Text Representation

Before the text is ready to be transformed into a representation that the machine can understand, the text should be preprocessed. All text is transformed to lowercase letters before applying stop-word and punctuation elimination. Afterwards, each post is tokenized, meaning that each word/term is added to a list as a single element. On each element, lemmatization is performed, a well-known NLP technique that transforms each word into the lemma of the word meaning its dictionary form. When the text is preprocessed, it needs to be represented so that machines can understand it. There are different methods to do so, but the methods used in the experiment are presented in this subsection.

n-grams

In the *n*-gram model, each token is called a gram. The *n*-gram model is defined as a first-order Markov Chain [Russel and Norvig, 2016, p.861]. The number n represents how many neighbor words are included in a term. It allows for more than one word as an entry.

Bag-of-Words

For training models for classification, the Bag-of-Words (BoW) model is popular to use when the frequency of a term is of interest [Singh, 2019]. The method creates a ”bag” that represents a document in the corpus. Suppose a corpus consists of 1000 documents, then the BoW method would give you 1000 bags. A bag can be represented as an ordinary list. Each entry in the bag represents a term likely given by a dictionary of the corpus. The bag represents the document as a list with the frequency of each term [Bengfort et al., 2018, p.56]. The list will often be sparse because a document will likely not contain all the terms chosen to represent.

TF-IDF

TF-IDF is a statistical method that calculates the importance of tokens in a document related to a corpus. The model investigates how often a token appears in a document and then calculates the frequency of each token in the corpus. TF_i is the percentage of token i of the total of tokens in the document. IDF_i represents how much information the token i gives in the specific document. It is based on how often i appears in the corpus. It is expected that words that appear a few times give more information than tokens appearing a lot [Bengfort et al., 2018, p.62]. Equation 3.1 shows the formula for calculating the weight of a term using TF-IDF.

$$tf - idf(t,d) = tf_{t,d} \times \log(\frac{N}{df_t}) \tag{3.1}$$

where t=term, d=document, N=number of documents in the corpus.

Word2Vec

Word2vec is a word embedding technique that wants to include the context of a word. It represents each word as a vector and aims to give similar or related words similar vectors. The model can be built on two different algorithms, namely, skip grams[1] and continuous BoW. The developer can choose how many words before and after a word to consider when calculating the word's representation. The probability does not pay attention to how far from the center word the context word is placed. Word2vec aims to give similar terms, in a similar context, similar representations [Bengfort et al., 2018, p.66].

Several parameters can be tuned in the word2vec model using Gensim, which will be explained in section 3.6. The two parameters tuned for the word2vec model are listed below. The training is an unsupervised process and does not have a good way to evaluate the results.

- *min_count:* Chooses how many times a word has to appear in the corpus to be considered in the model.
- *vector_size:* Is the dimension of the space that the word is mapped to. The larger it is, the more training time the program needs, but the model should deliver more accurate results. The value is reasonable when between tens and hundreds.

[1] Skip grams are not used in the project and therefore not further explained.

3.3 Sentiment Analysis (SA)

SA is a sub-field of AI that uses NLP and ML techniques. It is concerned with extracting opinions by determining whether a text is neutral, positive, or negative [Lexalytics, 2021]. It is solved by breaking the documents into segments that are given a weight representing the sentiment using a sentiment lexicon. Words like "bad", "nasty", and "kill" are given negative weights, while "excellent", "friendly", and "favorite" are given positive sentiment value. An obstacle with SA is that the sentiment value may vary depending on the use area and context.

The dataset used as a neutral set in the experiment is a dataset that has been classified using sentiment analysis, where each tweet is classified as either positive or negative. The dataset will be presented in chapter 5.

3.4 Machine Learning (ML)

ML is, like NLP, a sub-field of AI. ML algorithms are used to learn by automatically training on existing data or giving rewards and punishments to the agent. The goal is that the ML algorithm builds a real word model that can make decisions and predictions without the help of humans [Bengfort et al., 2018, p. xii]. After the natural language text has gone through the process of NLP by being preprocessed and represented, it is passed to the ML algorithm. Four common approaches of ML are *Supervised*, *Unsupervised*, *Semi-supervised* and *Reinforcement* learning approaches.

Supervised learning is a technique that uses known data to train a model. When training the model, the results of the input should be known. Hence the parameters of the model can be tuned along the way. For supervised algorithms, there exist two approaches, either classification or regression. Classification labels the output as a category, used for discrete values, whereas regression is used when working with continuous values. **Unsupervised** techniques do not use predefined data to tune the model. Instead, it has to analyze the input data to find patterns. The result of an unsupervised algorithm is the grouping, or clustering, of data. **Semi-supervised** techniques combine supervised and unsupervised learning. It uses a large amount of unlabeled data and a smaller set of labeled data. It can be beneficial to use only with limited data available. **Reinforcement** learning trains the model by defining the ideal outcome. The rewards and punishments are given based on the deviation from the ideal solution to tune the parameters. Despite that reinforcement learning has had increasing success in the later years in many fields, it is not the case for NLP [Uc-Cetina et al., 2021].

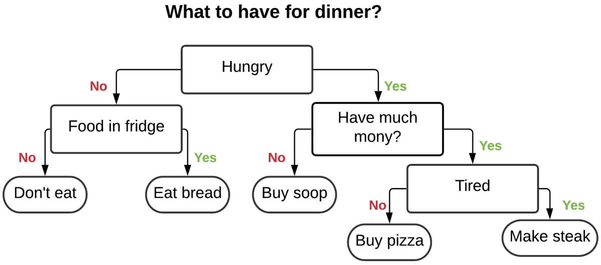


Figure 3.1: Deciding today's dinner using a DT

3.4.1 Supervised Machine Learning Algorithms

For all of the following algorithms, there exist different variations. In this subsection, the basic versions are the only ones explained.

Naïve Bayes (NB)

Naïve Bayes Classifier is a probability-based method that is easy to both implement and use. It is based on Bayes theorem (Equation 3.2). It is called naïve since it assumes that every feature is independent of the others. That assumption rarely holds in the real world but has performed well despite this. If the assumption holds, it converges faster than most other methods. NB returns the probability of belonging to a category [Brownlee, 2016].

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)} \tag{3.2}$$

A=Expected classification, C=Evidence (observed features)

Support Vector Machines (SVM)

SVM is a popular geometric-based method to try when solving a classification problem and can also be applied for regression problems. It generalizes well. Each data unit is plotted in a n-dimentional space as a single point. Then SVM creates a hyper-plane that separates the data points into different categories. The aim is to find the maximum margin, meaning the maximum distance from the data points of the different classes. SVM can be used in high-dimensional spaces. This is an advantage since the data rarely is linearly separable. The method is said to combine the best of non-parametric and parametric methods because it can represent complex functions without overfitting [Russel and Norvig, 2016, p.744-748].

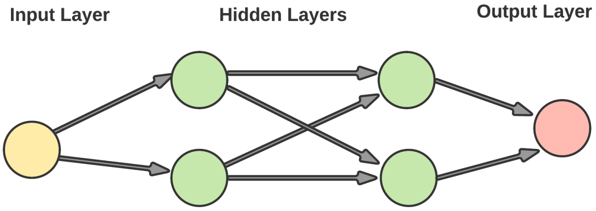


Figure 3.2: A simple illustration of an ANN taken from the specialization project.

Decision Trees (DT)

DTs can be used for classification, even multi-label classification, but can also be used for regression. When used for regression, the trees are often called regression trees [Yadav, 2018]. Such models are often fast to train and easy to understand [Mohri et al., 2012, p.195]. The DT has a tree structure where each feature is represented as a node, and every leaf node is a decision or category. Due to its resistance to outliners, it requires little preprocessing but is prone to overfitting and creating biases if not equally distributed data for each class. An example of a DT is shown in Figure 3.1, where it is used to decide what to have for dinner.

Random Forest (RF)

The Random Forest technique is based on DTs and creates a "forest" of DTs [Mbaabu, 2020]. The main difference from DTs is randomly choosing of root nodes and segregating nodes in RF. It uses bagging, which is to use multiple samples of training data instead of just one. Each of the DTs in the "forest" delivers a classification, and the category classified by most sub-trees is chosen as the category predicted by the RF [Cutler et al., 2011].

3.4.2 Clustering

Clustering is a method used in unsupervised learning approaches. It aims to cluster, or group, the data by finding patterns or structures in the training data [Alashwal et al., 2019]. In the retrieved literature, few used clustering.

3.4.3 Deep Learning (DL)

DL is a sub-field of ML utilizing the biological brain's structure and function to create artificial neural networks (ANNs). Most DL algorithms are supervised, but such algorithms can also be unsupervised, semi-supervised, or reinforcement methods. DL allows for less preprocessing than the other mentioned approaches, and it has the ability to extract features automatically. Hence, DL requires less supervision from humans [IBM Cloud Education, 2020a]. An ANN consists of multiple layers, one input layer, one output layer, and a number of hidden layers [IBM Cloud Education, 2020b]. Figure 3.2 is included to show the relations between the layers, but it is a simple illustration. All the layers can have more nodes, and the number of hidden layers could be higher. Each node in the network simulates a biological neuron that is connected to other artificial neurons with an associated weight and threshold. If the input value to the artificial neuron passes the threshold value, it is activated and delivers an output value. The networks learn by adjusting the neurons' weights by monitoring the error rate observed.

In a neural network model, multiple parameters can be tuned to improve the model's accuracy. Size, activation functions, optimizers, loss functions, and learning rates were modified during the developing phase.

Sizes

The two sizes that were modified in the experiments were the layer units and the batch sizes.

- **Layer Units:** The units defined for a layer is the size that the output from the layer should have.

- **Batch Size:** A batch size is chosen when training a model or using a model for prediction. It is the data size used in each iteration in one epoch.

Activation Functions

The activation function determines how the input should be transformed into the output given the weighted sum calculated in the layer. It is common for the hidden layers to have the same function, while the output layer's activation function is usually another than the rest of the layers. For all models, ReLU was used in all layers except the output layer where Softmax was used.

- **ReLu:** The ReLu (Rectified linear activation) function is considered the most common one for hidden layers due to its simplicity and efficiency. The function returns the value as long as it has a positive value [Brownlee, 2019].

- **Softmax:** The Softmax function is a standard way of transforming numbers into probability distributions. For multi-class classification the softmax activation function is considered the one to choose [Koech, 2020].

Optimizers

The optimizer is used when the model is compiling and when it is training. The task of the optimizer is to minimize the loss.

- **Adam:** Adam is an adaptive optimizer meaning that the learning rate does not necessarily need to be tuned automatically. Generally, Adam is considered to be the best option [Kingma and Ba, 2014].
- **Stochastic Gradient Decent (SGD):** For SGD, it has to be defined a learning rate or a scheme for calculating the learning rate. It calculates the gradient. Sometimes, it is wise to test SGD together with a learning rate schedule because it may give better results than Adam if training enough [Giordano, 2020].

Loss

A loss function is implemented to show the model how much it should look for a better solution by minimizing the loss during training [Keras, 2022]. It had to be chosen a loss function that handles multi-label classification.

- **Sparse Categorical Crossentropy:** The function compares the labels and the prediction by computing the cross-entropy loss [TensorFlow, 2022].

Learning Rate

The learning rate defines how often the model updates its weights during the training. A low learning rate may result in the need for a larger number of training epochs than higher learning rates. The drawback of higher learning rates is that it can make the model satisfied with its prediction before it should.

- **Static:** It could be defined a learning rate that should apply throughout the entire training process.
- **Schedule:** It was implemented a schedule called *ExponentialDecay* that decreases the learning rate during the training starting with an initial learning rate and a decay rate.

3.5 Evaluation

Table 3.1 is taken from the specialization project and demonstrates four different categories to divide the predictions into. It is desirable to maximize the true positive and true negative predictions. A common evaluation method is accuracy which is calculated using Equation 3.3 and tells how many correct predictions

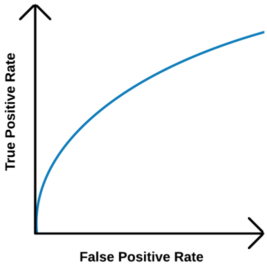


Figure 3.3: The ROC curve

the model made. Precision (Equation 3.4), recall (Equation 3.5) and F1 score (Equation 3.6) are three other metrics used to evaluate a model. Accuracy is a good metric to use when dealing with a balanced dataset. If it is desirable to reduce the FP, the precision metric should be used and tried to improve, while recall should be used if the goal is to minimize the FN. F1 score is used when both tasks are important and therefore combines precision and recall in its evaluation [Santos, 2020].

Area Under the Curve (AUC) is a metric telling the model's ability to determine the difference between classes by calculating the area under the curve. The curve is the Receiver Operating Characteristic (ROC) curve. The ROC curve can be calculated by using TP rate, the same as recall, and FP rate as shown in Figure 3.3 [Bradley, 1997].

	Predicted Negative	Predicted Positive
Is Negative	True Negative (TN)	False Positive (FP)
Is Positive	False Negative (FN)	True Positive (TP)

Table 3.1: Confusion Matrix

$$Accuracy = \frac{TP + TN}{All\ Predictions} \tag{3.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.5}$$

$$F1\text{-}score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.6}$$

3.6 Technical Tools

Different tools were used during the experiments not to have to implement them all from scratch.

- **Garc:** Garc is a interface for Gab API[2]. It helps collect JSON objects from Gab. The interface can be used to collect a JSON file for each user where each post is represented as an instance containing metadata about the post and the account. An example of how a post looks in the JSON object can be found in the appendix in Listing 1.

- **Pandas:** The posts retrieved using Garc can be extracted from all the retrieved JSON objects and represented as a pandas series. Pandas is a tool made on Python to help analyze data and manipulate it[3]. It has the advantage that it is fast, flexible and easy for the programmer to use.

- **Tweepy:** A easy library for accessing the Twitter API is called Tweepy[4]. As will be mentioned in subsection 5.1.2, Twitter was not scraped to collect any datasets for the experiments since it was found suitable datasets online. Hence the library was only used to check if Gab users were on Twitter. The status of a username could be not existing, suspended or banned users.

- **NLTK:** The Natural Language Toolkit (NLTK) is a leading platform to use when developing Python programs for natural language text[5]. It delivers tools for easy preprocessing of text such as tokenizer, lemmatization and help for removing stop words.

- **Num2Words:** The library can convert numbers in digits to numbers as text[6]. Num2Words should be used when numbers are considered useful since, if not transformed to text, the digits would be removed during preprocessing. Additionally, it transforms numbers of different formats to be represented in the same matter, making it possible to analyze the frequencies of numbers.

- **Gensim:** Gensim is a library for Python for topic modelling popular to use for NLP[7]. Gensim can be used for preprocessing and to build Word2Vec models.

[2] Git documentation for Garc: `https://github.com/ChrisStevens/garc` (Accessed: 02.05.22).

[3] Pandas documentation: `https://pandas.pydata.org` (Accessed: 02.05.22).

[4] Tweepy documentation: `https://docs.tweepy.org/en/stable/` (Accessed: 02.05.22).

[5] NLTK documentation: `https://www.nltk.org` (Accessed: 02.05.22).

[6] Num2Words documentation: `https://pypi.org/project/num2words/` (Accessed: 02.05.22).

[7] Gensim documentation: `https://pypi.org/project/num2words/` (Accessed: 02.05.22).

- **Sklearn:** The Sklearn library is popularly used for predictive data analysis[8]. From the library the three following tools were used:

    1. *CountVectorizer* that transform a text corpus to a matrix of counts.
    2. *TfidfTransformer* that helps calculating the TF-IDF scores.
    3. *train_test_split* splits the dataset into appropriate training and test sets.

- **Tensorflow:** TensorFlow is developed by Google as a end-to-end platform for ML[9]. The library delivers high-level API for building and training ML models, but Keras is considered more user friendly than TensorFlow.

- **Keras:** Keras is an API for building and training artificial neural network models[10]. It is developed on TensorFlow and in Python to be easy to use. The Keras works as an interface for TensorFlow.

[8] Sklearn documentation: `sklearntraintestsplit` (Accessed: 02.05.22).

[9] TensorFlow documentation: `https://www.tensorflow.org` (Accessed: 02.05.22).

[10] Keras documentation: `https://keras.io` (Accessed: 02.05.22).

Chapter 4

Previous Work

In this chapter, work performed by others in related studies will be presented. Most of the findings were gathered during the structured literature review but are included in the master's thesis because of its importance for understanding the need of the project work. Section 4.2 is taken from the specialization project. Before the findings are presented, a short introduction to the method used to retrieve the relevant literature is given.

4.1 Structured Literature Review Protocol

Structured Literature Review (SLR) is a method used to retrieve relevant literature given some research questions formulated as queries. The variant of SLR used in the specialization project was proposed by Kofod-Petersen meant to be suitable for the field of computer science [Kofod-Petersen, 2018]. SLR does not guarantee that the delivered results are relevant, but it excludes the majority of irrelevant literature. The two research questions from the specialization project that were used to retrieve the literature that the findings and previous work are based on are:

1: *What are the most promising approaches to predict who will be radicalised?*

2: *What are the issues defining people vulnerable to extremism?*

The query in Equation 4.1 was formulated to retrieve the literature that could assist in answering the overall goal and the research questions.

$$
\begin{gathered}
(Extremism \lor Radicalism \lor Terrorism \lor Right-Wing)\land \\
(Social\ Media \lor Twitter \lor Facebook \lor Gab)\land \\
(Prevention \lor Identifying \lor Detection)\land \\
(Artificial\ Intelligence \lor Machine\ Learning\lor \\
Prediction \lor NaturalLanguageProcessing)
\end{gathered} \tag{4.1}
$$

In addition to using SLR, other literature referenced in the retrieved literature and recommended literature by the supervisor were included in the review. A detailed description of the retrieval process is included in the appendix in section C. The chosen literature is described in section D.

4.2 Related Work

The whole section is copied from the specialization project with only minor changes. It was considered the best option since it is the background for the master's thesis.

The related work shows that most existing studies are concerned with detecting extreme content or users. A few articles address different approaches for predicting people in the radicalization process on social media or people vulnerable to social media extremism. The minority of the literature utilized psychological or social radicalization theories, but this was not explicitly searched for in the query, so not surprising. One aim of the SLR was to retrieve literature that discussed the prediction of radicalization using NLP and ML. Despite this, several articles concerned with network analysis were returned, lacking the use of NLP in their studies. It was considered if this literature should be removed. However, most of them were kept because their findings could inspire future work since the approaches may be adjustable to fit the processing of natural language texts. All literature in this section is related to radicalization and social media.

4.2.1 Predicting Radicalization

A literature review done by Gaikwad et al. [2021] did not deliver any valuable results for future work. However, it gave a good overview of existing solutions, relevant datasets, and approaches to focus on reading the rest of the literature. It shed light on different shortcomings of the existing research and proposed that future work should deliver solutions that are not so ideology dependent.

Asif et al. [2020] studied using sentiment analysis to classify comments and posts on news pages on Facebook into four degrees of extreme content. To do so, they used TF-IDF and trained the model by applying Multinomial NB and SVM. The authors conducted a lexicon with terms containing a belonging sentiment value to each term due to the lack of multilingual lexicons for extremism. The SVM delivered the best results with an accuracy of 82% but did not yield satisfying results for the second-highest degree of extremism. López-Sáncez et al. [2018] proposed an approach partly automated for the process of identifying influential users and monitoring their interaction to retrieve and manually accept or reject users at risk of radicalization. It was desirable to estimate the overall vulnerability of being radicalized. Only a small case study was performed, but future work would utilize both text and network interactions.

Psychological models have been used as inspiration in multiple cases of studies on radicalization on social media, such as for Fernandez et al. [2018] who used the "roots of radicalization" that is presented in subsection 4.2.3. The study aimed to use this model to predict the risk of someone being radicalized rather than identify if someone is radicalized. They found few studies cornered with predicting radicalization. Those who did focus on a few features, not the "whole" package as Fernandez et al. wanted when using the "roots of radicalization model". In subsection 4.2.3, it was explained how the factors of radicalization were adapted to the social media environment.

A methodology was proposed by Al-Saggaf [2018] to recognize when youths start to express radical beliefs until they are radicalized. The grievances to win the sympathy of young people expressed by extreme groups were investigated. Further, they expressed their aim of using neurolinguistics[1] to distinguish between the radicals and non-radicals and recognize when youths are close to the activation point. Rowe and Saif [2016] used a computational approach to find when a user reaches the radicalization point. They aimed to figure out the behavior of pre-radicalized individuals on social media and use it to implement measures. The study found that the feeling of a relationship with like-minded is a drive for radicalization, and the signs to monitor users were divided into the three dimensions: *lexical*, *sharing* and *interactions*. It was unveiled that sharing was a better indicator than the lexical dimension. Another approach used for predicting radicalization was studied by Tundis et al. [2018] by calculating the suspiciousness of users on social media by text analysis. The study started to detect users that possibly support organized crime or terrorist networks. Delta term frequency, n-grams, and BoW were used to calculate the suspiciousness and resulted in an accuracy of 79%. ML was not used to train a model, but it was expressed a desire to do it in further work.

[1] Neurolinguistics is the study of how language is present in the brain. `https://mitpress.mit.edu/books/neurolinguistics` (Accessed: 04.12.21)

Preventing acts that threaten national security using social media is of great importance for nations. Cardenas et al. [2018] propose that inspecting users prior to, during, and after threatening acts can help to detect users that could be near the activation point [Cardenas et al., 2018]. A lexicon-based approach was used and trained. The model was tested and trained with Gradient Boost Machine Model (GBM), RF, and DL, where DL delivered the best accuracy of 94.6% for AUC. Expressions related to defense, health, and government were shown to be related to the likelihood of posing a threat to national security. Ferrara et al. [2016] aimed to detect extreme users, predict the likelihood that regular users will adapt to extreme content and acquire extreme beliefs and behavior [Ferrara et al., 2016]. In addition, predicting if users will answer extreme users contacting them. The features used were related to either metadata, network statistics, or timing features, not utilizing NLP techniques. Using RF and LR, the three tasks were trained both in real-time and time-independent. The best AUC for the time-independent tasks ranged from 72% to 93%, and the most significant feature was discovered to be the ratio of tweets/retweets.

The study by Kursuncu et al. [2019] aimed to deliver a framework to understand the radicalization process to be able to implement counter-programming. Another focus was to minimize the discrimination biases of false positives for non-extremists. An account was represented in three dimensions, namely *religion*, *ideology* and *hate*. The experiments showed that a combination of all dimensions outperformed competitive baselines. NB and RF were tested, and using RF performed the best, delivering an AUC of 93%. Beheshti et al. [2020] built their research on cognitive science by using the golden standards for personality, behavior, and attitude to build a knowledge base. The study was generalized, meaning that the approach can be applied to other extreme areas than radicalization. They aimed to build a dashboard that shows a pipeline of the users' behaviors. Content, context, and activity were analyzed and should be displayed in the dashboard. Arya et al. [2019] also proposed a generalizable framework. The graph-based approach aims to predict future interaction in forums with extreme beliefs using multimedia features.

Lara-Cabrera et al. [2019] wanted to estimate the risk of radicalization for individuals using different indicators, defined by psychologist experts, for behavior expressible in social media. They focused on the individual's frustration, level of introversion, perception of discrimination for being Muslim, negative ideas about Western society, and finally holding positive ideas of Jihadism. The study showed that it was more likely for radicalized users than for ordinary users to use swearing and negative words. The posts also tended to be longer and express negative feelings about the Western society but positive ideas about Jihadism.

The three latter indicators were concluded to be the most specific for extremism. An approach relying only on keywords could miss essential information if it does not contain enough words, variants, etc. [Barhamgi et al., 2018]. They wanted to rely on semantics, not keywords, by introducing concept annotation.

4.2.2 Analyzing and Identifying Extremism

Kursuncu et al. [2019] proposed a framework that is concerned with extremism in general and that aims to recognize rebel users such as extremists. It uses both textual features and several account features in a graph-based approach. Rehman et al. [2020] delivered results suggesting that radical features should be included in addition to religious features for radicalization detection. It was investigated if radicals use more violent and bad words, but the results were not convincing enough to conclude. An approach by Xu et al. [2017] consisted of creating a structure of supporters of extreme organizations and calculating the estimate for future terrorism activity. Mussiraliyeva et al. [2021] presented software to identify extreme users, communities, and resources spreading radical content. The characteristics of such content were of interest and the ability to compute the risk that the members of extreme groups pose. In a previous study by Mussiraliyeva et al. [2020], Word2Vec and TF-IDF were used with RF and Gradient Boosting, where all the combinations delivered results right below 90%. The aim was to classify content from VKontakt as either expressing extreme behavior or not.

Statistics show that fewer than 1% of radicals turn into terrorists according to [Wolfowicz et al., 2021]. The article expressed their concern that using only text-based features to identify violent extremists is limited due to the limit of sentiment it discovers and the difficulty of understanding the context and the rate of false positives. The study wanted to find social theory features for classifying radicals as non-violent or violent. They retrieved user information 100 days before an attack to recognize essential features implying a violent extremist. It was discovered that terrorists posted less content but instead shared existing content prior to the attack. Munk [2017] found that the existing approaches for detecting terrorists return 100,000 false positives for each true terrorist[2]. The study by Nouh et al. [2019] wanted to identify radical content on social media by analyzing different signals such as textual, psychological, and behavior. It showed that the psychological signals were the most distinguishable but that radicals, in general, were different from ordinary users in all three areas. One model analyzes the importance of a word, the second model the semantics of the language, and the third is a psychological model. Using TF-IDF and Word2Vec in combination

[2] The article was written in 2017, meaning that the existing approaches do not include what has been done after.

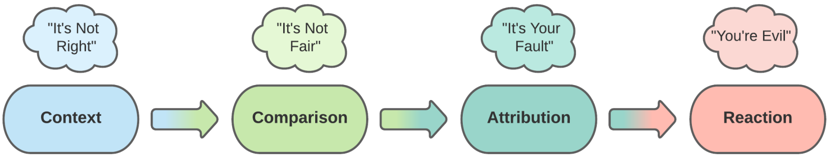


Figure 4.1: Randy Borum's Process of Ideological Development.

with different ML techniques, it was discovered that RF and NN performed best.

The study by Udanor and Anyanwu [2019] investigated the percentage of hate speech on social media and how much these platforms tolerate that. To detect hate and extremism promoted on Twitter, Agarwal and Sureka [2015] applies SVM and k-Nearest Neighbors (kNN) using multiple linguistic features. Hashtags were used to conduct a dataset for training. It was discovered that SVM performed the best with an accuracy of 97% and that religious and war-related terms are frequently used in extreme content. For the SVM model, slang and question marks were significant. Hashtags were the only textual feature used in the study by Benigni et al. [2017] to give insight into social media extreme networks. The study expressed the need to understand who the vulnerable users are to implement counter-measures. It was discovered that passive supporters were targets for the extreme groups to recruit to become active extremists. Unlike the rest of the literature, Aleroud et al. [2020] tried to use a term augmentation technique which resulted in higher precision, recall, and F-score than the usual sentiment analysis using TF-IDF.

4.2.3 Radicalization Process

Different radicalization models exist to explain the process of people having ”normal” opinions starting to acquire radical beliefs and possibly becoming radicalized. It is essential to understand the process to recognize the early stages and hopefully stop the rest of the radicalization process. This section will be helpful in getting familiar with typical warnings in behavior and theories about considering when a person has passed the activation point. The activation point is where a person passes the radicalization process and becomes an extremist. This section presents three theoretical models for the radicalization process and online radicalization.

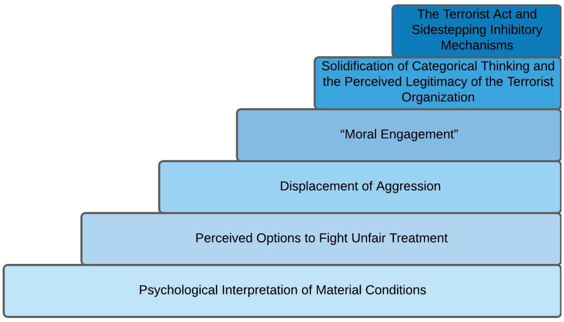

The Terrorist Act and Sidestepping Inhibitory Mechanisms
Solidification of Categorical Thinking and the Perceived Legitimacy of the Terrorist Organization
“Moral Engagement”
Displacement of Aggression
Perceived Options to Fight Unfair Treatment
Psychological Interpretation of Material Conditions

Figure 4.2: The Staircase to Terrorism.

Four-Stage Model

The four-stage model was presented by Borum [2003]. The aim of the model is to be a tool for analyzing the behaviours and activities of individuals or groups/organizations that is prone to radicalization. The four stages are *Context*, *Comparison*, *Attribution* and *Reaction* like in figure 4.1 [Borum, 2003].

In the first stage, the individual, or group, starts recognizing experiences they feel are unfair and unsatisfying. It can be multiple reasons for the feeling, such as dissatisfaction with their life situation. Borum suggests that the result is that the people in the context stage start to think ”it is not right”. Moving to the next stage, the individual or group feels that the statement is not applying to everyone. Hence, a natural thought would be ”it is not fair”, leading to resentment. In the next stage, the people start seeking someone to blame for the inequality saying, ”it is your fault”. Reaching this point, the people are in the process of being indoctrinated. Typically the government, one race, or one religion is blamed. At the last stage, they blame the target, saying, ”you are evil”. Aggression is built against the ”evil” part and dehumanized by the individual or group. Now they have turned into extremists.

Staircase to Terrorism

Moghaddam [2005] presented a staircase model to explain terrorism and the radicalization process. The model consists of a ground floor and five floors above, where the last one represents when people perform acts of violence. Figure 4.2 shows Moghaddam’s model and the name of each floor. Every floor holds one behavior categorized by a psychological process.

The majority of the world’s population never leaves the ground floor where they

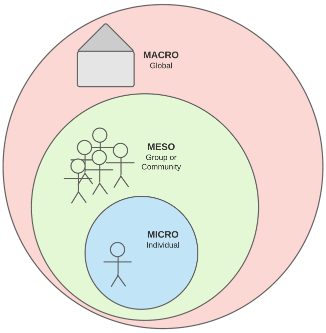


Figure 4.3: Roots of Radicalisation.

appreciate the experienced fairness. The people who are climbing to the first floor experience injustice. On the first floor, the people try to find a solution to the perceived unfair treatment. If not found, they step up to the second floor. Since the situation could not be solved on the first floor, the people on the second floor start to experience frustration and anger. Here they try to find a target to hold responsible, like the government or a race. If they find one, they move to the third floor. On the third floor, violent organizations can start to show interest in recruiting them. The people on the third floor are invited into groups with familiar enemies and similar goals as themselves. People that find it exciting move to the next floor. When arriving at the fourth floor, the probability of withdrawal is close to zero. Then the thought of "us versus them" arises. The individuals isolate themselves from family and friends and put the organization above others. At this stage, they wait for an opportunity to move to the last floor. The fifth floor is where the now extremist are ready to perform an act of terror.

Using this model, prevention needs to be done before entering floor three. Preferably and most accessible, it should be done on the ground floor.

Roots of Radicalisation

The Roots of Radicalization model digs into the different reasons, roots, leading to radicalization. The three different roots are *Micro* (Individual), *Meso* (Group) and *Macro* (Global) which each describes different aspects of life as Figure 4.3 presents. Dr Schmid has described these aspects as causes of radicalization, possibly leading to terrorism [Schmid, 2013, p.4].

- **Micro:** Individual is experiencing negative feelings related to factors on an individual level. The feelings can be related to deprivation, injustice, and discrimination.

- **Meso:** At the group level, the individual finds support from groups and organizations. The groups can offer a feeling of belonging and confirm the individual's ideas. Comparing themselves to other groups to show injustice promotes the "us versus them"-feeling.
- **Macro:** The global level represents the affected factors of a government or a society. It could be in its nation or another country. Factors can include political parties' opinions and attitudes towards immigrants or minorities.

According to Fernandez et al. [2018], it is easy to implement in the digital world. At a micro-level, the internet can help the individual with easy access to content and facilitate self-radicalization. At the meso-level, the extreme groups can reach a larger audience on the internet without physically meeting. Hence, it is easier to meet people that support your ideas. Lastly, the internet allows for fake news, propaganda, and radical content in the vulnerable monitors. This is how the macro-level can be transferred to the internet.

Online Radicalization

Behr et al. [2013] states that self-radicalization is allowed for acceleration if the individuals are allowed to uninterrupted communicate with like-minded people. The interactions are more available when social media can be used as a communication tool than if they had to meet in an old-fashioned way. Social media is also used by extremists and extreme groups to recruit and radicalize [International Association of Chiefs of Police, 2014].

4.2.4 Counter-Radicalization

After 9/11, the focus on radicalization increased both on the government level and in the population. The UK has implemented a counter-terrorism strategy called *CONTEST* which aims to *Prevent*, *Pursue*, *Protect* and *Prepare* to protect the population of UK [Secretary of State for the Home Department, 2018]. Skleparis and Knudsen [2020] addresses the differences between counter-radicalization in the UK and Greece. The UK has been a pioneer in the field, while Greece has been late. Unlike the UK, Greece has no legislation for preventing radicalization, but they punish acts of terror. It is shown that radicalization knowledge has been localized locally, not internationally, which might be undesirable. Despite that knowledge should be shared, it is different how to counter radicalization in different locations. The article introduces three "truths" about radicalization; radicalization can lead to terrorism, the process can be stopped or reversed at an early stage, and radicalization is measurable. The England and Wales' Risk Guidance (ERG22+)[3] and Vulnerable Assessment Framework (VAF)[4] are tools focusing on risk indicators, 22 which are similar for them both. The indicators are used to assess the risk of non-criminal individuals. Some disadvantages of using these tools are that they do not capture broad political, societal context or content, or type of radicalization. It raises the question of how to assess the risk of radicalization when people with different occupations and experiences evaluate it.

In Britain, the school is considered an institution that can build resilience towards extremism at a young age by teaching them the "British Values"[5] [Winter et al., 2021]. The school should report about students and staff possibly vulnerable to radicalization. Winter et al. [2021] performed interviews at two secondary schools by using the color-blindness concept as a theoretical framework. Color-blindness is when someone says they do not precept race or let it affect their decisions, meaning that they must not show any performance of racial bias. It was unveiled that the students often associated terrorism with Muslims. Additionally, the classrooms seemed to be a helpful arena for recognizing suspicious beliefs that might suggest vulnerability to radicalization.

[3] Is developed to assess the sentence of terror-related crimes.

[4] For local authorities to assess the possible risk of radicalization of individuals reported in the *Prevent* stage.

[5] British Values: "democracy, the rule of law, individual liberty, mutual respect, and tolerance of different faiths and beliefs". `https://wncqtlp.wixsite.com/prevent/british-values` /Accessed: 05.11.21)

Chapter 5

Experiments

When having understood the background theory and related work, the next step is to introduce the plan for the experiments and how the experiments were performed. For all experiments, python was used as the programming language, and Visual Studio Code[1] *was used as the code editor.*

5.1 Experimental Plan

This section gives an overview of the experiment and the plan for performing the experiments. For each of the choices made in the planning phase, it is argued why they were made. In addition, it is included argumentation for the importance and necessity of the experiments justified by previous work of others and the work in the specialization project. Figure 5.1 shows the five experiments that were performed. First, it had to be gathered data, then analyze the text to look for patterns, followed by preprocessing the text to apply NLP and ML techniques.

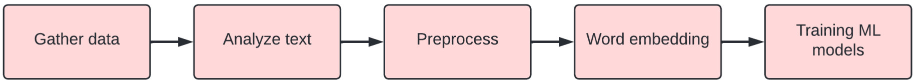


Figure 5.1: Plan for the experiments.

Kennedy et al. [2018] expresses the need to be careful when studying the field of hate speech due to the psychological effects it can have on the individual to be exposed to hateful expressions frequently. Hence it was decided early in the

[1] Homepage for Visual Studio Code: `https://code.visualstudio.com` (Accessed: 01.05.22)

semester that reading of content published by extremists should be minimized to what was needed for delivering good results and instead focus on the technical issues.

5.1.1 Choice of Social Media Platforms

Facebook, Twitter, Gab, VKontact and Stormfront were all introduced in section 2.3 because they were considered to be suitable options to gather data from. The final choice of platforms fell on Twitter and Gab due to the character limits and their similarities. It turned out that Gab had increased its limit from 300 to 3,000, making the character gap between Twitter and Gab larger. The change was discovered after the data had been gathered, and therefore it was decided to keep the dataset but remove the posts with more than 300 characters. Two platforms were chosen instead of one because finding non-radical or non-extreme content turned out to be difficult on Gab. Therefore it was decided to find neutral data from Twitter and radical and extreme data from Gab. The founder of Gab has described Gab and its purpose in his gab feed as[2]:

> *Gab is a First Amendment company which means we tolerate ”offensive” but legal speech.*
>
> *We believe that a moderation policy which adheres to the First Amendment, thereby permitting offensive content to rise to the surface, is a valuable and necessary utility to society.*
>
> *It allows unorthodox but correct views, such as the Wuhan Lab Origin Theory that was banned on Twitter and YouTube but permitted on Gab, to propagate.*
>
> *It allows hateful ideas, such as anti-White CRT, to be exposed and subject to scrutiny and challenge. It also allows Americans, and others around the world, to enjoy the full measure of their human right to speak freely online.*
>
> *Supporting the mission of freedom online means having the stomach to accept that people will say ”edgy and offensive” things.*

[2] The URL for the gab post: `https://gab.com/a/posts/106508069363422579` (Accessed: 01.06.22)

Below, a few gabs are listed to show content allowed to be posted on Gab and not removed. The gabs would probably be removed from Twitter and Facebook due to the hateful and racist content.

- *The best way to stop White genocide and White replacement, both of which are demonstrably and undeniably happening, is to get married to a White woman and have a lot of White babies.*
- *Reminder: Google is jewish owned and hates white people. Google has manipulated its search algorithms in a way that shows their utter contempt of white people.*
- *Message to black people: Stop killing white people. Thanks!*

5.1.2 Datasets

Asif et al. [2020] performed an experiment classifying radicals into four groups with different levels of radical using ML. The experiment resulted in an idea for the master's thesis to gather three different categories of datasets and use them to classify posts into three. The three chosen categories were decided to be *neutral*, *radical* and *extreme*. The definitions of radical and extreme for this project are defined in subsection 1.4.1, and the approach for gathering the tree datasets is presented in section 5.2. The neutral dataset was found online, so there was no need for scraping Twitter, but the dataset had to be preprocessed and analyzed. Subsection 5.2.1 describes the neutral dataset and where it can be found.

Novel Dataset

None of the retrieved literature found using the SLR in section 4.2 used data from Gab or other publicly available right-wing datasets. Neither was a suitable right-wing dataset found when searching for it on open sources. A dataset containing extreme right-wing content was necessary to perform the planned experiment. Hence it was necessary to gather a novel right-wing dataset.

5.1.3 Text analysis

Lara-Cabrera et al. [2019] delivered results stating that radicalized pro-ISIS users tend to use more swearing and negative words than neutral users. It was decided to investigate if the most frequently used terms in the right-wing datasets gathered from Gab contained more negative words than the datasets from Twitter. Rehman et al. [2020] investigated if the used words by radicals were more violent and bad than words used by non-radical users. However, unfortunately, the results were not convincing enough to conclude. Masood and Abbasi [2021] suggests that the hashtag use of radicals differs from neutral users since they often use the platform to promote their cause and therefore use more hashtags and mentions than others. Based on these suspicions, the hashtag and mention use were investigated for each dataset. In the roots of radicalization model, described in section 4.2.3, URLs are used as a feature to analyze the macro aspects of the process. Hence, a brief analysis of URL usage was decided to be performed. Additionally, the length of the posts was calculated and compared, which is motivated in subsection 5.3.1. The process for text analysis is explained in section 5.3.

5.1.4 Preprocessing

Before presenting the plan for word embedding and ML, the text needed to be preprocessed as described in subsection 3.2.1. The approach used for each of the datasets will be described when explaining the experimental setup in section 5.4.

5.1.5 Word Embeddings

After analyzing the text to map patterns in each dataset, it was desirable to use NLP techniques to represent the data in other ways to look at the importance of the different words and the similarities between words in the datasets. Therefore, the plan was to apply TF-IDF and word2vec. It would be helpful to be able to calculate the degree of how radical someone's posts are, which resulted in a plan to investigate the possibility of using TF-IDF scores to give a post or user a score describing how radical it is.

5.1.6 Machine Learning

Asif et al. [2020] delivered promising results facing a similar task as this project of classifying short natural language texts into four categories. When training their model, they used NB and SVM. Since the task was similar to this master's thesis, it sounded interesting to try using ANNs for this multi-label classification problem. Cardenas et al. [2018] tried different ML methods to predict users near the activation point and received the best AUC accuracy for DL. Nouh et al. [2019] found that when combining Word2Vec and TF-IDF with ML approaches, RF and DL delivered the best results. Multiple previous research suggested the promising use of DL, which resulted in the chosen ML approach for this experiment to be DL.

Using NB and RF, Kursuncu et al. [2019] focused on reducing FP of non-extremism, which delivered an AUC of 93%. Reducing the number of actual neutral posts classified as extreme or radical was a motivation for the experiment. It was interesting to investigate whether using DL would be promising for further development. The need for developing models that reduce the number of FP is shown by Munk [2017] which presents that for every actual terrorist, the existing models return 100,000 FP users.

5.2 Data Gathering

A large part of the work of the master's thesis turned out to be the process of gathering data due to the lack of desirable available datasets. This section explains how the neutral dataset was found before it is given a detailed description of how the extreme and radical datasets were scraped from Gab. Lastly, the dataset used for analyzing the results the models gave is presented.

5.2.1 Neutral Data

While surfing on Gab to get to know the platform, it was experienced that most randomly found users expressed more radical content than expected for "normal" users. Hence it was considered difficult to gather a neutral dataset from Gab, and it was not found an open-source Gab dataset suitable to use as neutral data. Therefore it was investigated if there existed already conducted and free datasets online consisting of social media posts with a similar structure as gabs, preferably tweets. On Kaggle[3] it was found a dataset called "Sentiment140 dataset with 1.6 million tweets"[4] that contained tweets that were classified as either positive or negative. Since it is common for regular users to post both positive, neutral and negative content, it was considered useful to include both positive and negative tweets in the neutral dataset. An assumption was made that including normal negative tweets would decrease the rate of FP in the ML models. The file size was 238.8 MB, resulting in only using part of the file. About 320,000 posts classified as negative and 256,000 positive classified tweets were chosen. The final dataset was 83.1 MB and contained about 577,000 tweets when duplicates were removed.

5.2.2 Extreme Data

It was found a post on Gab posted by a known extremist that encouraged Gab users to comment *"Help find me find my frens"* if they were banned from Twitter. The post was posted in the latter part of 2021 and had over 3,000 likes, more than 1,500 comments and about 900 reposts. Using Garc, see section 3.6,

[3] The Kaggle website: `kaggle.com`

[4] The dataset is available here: `https://www.kaggle.com/kazanova/sentiment140` (Accessed: 15.02.22)

users that had commented the phrase were scraped from Gab. A JSON file was created for each user containing its posts created from 2005 until now, formatted as shown in the appendix in Listing 1. Tan et al. [2014] states that about 50% of social media users use the same username on different social media platforms. To verify if the collected users could be classified as extremists it was checked if the username was banned from Twitter. When searching for a user on Twitter, it can return that the user exists, does not exist, is suspended, or is banned. Using Tweepy, which was described in section 3.6, each of the usernames was checked. If the status code returned was 63, the user was suspended and met the requirements to be considered extreme in this experiment. Since Gab has increased its character limit from 300 to 3,000, only the gabs under 300 characters were added to the extreme dataset. Reposts and duplicates were removed, so the dataset only contained original posts. The final dataset contained almost 100,000 posts, and the file size was 41 MB. Most likely, not all of these posts would be classified as extreme, but to simplify the work, it was chosen to assume that all posts contained extreme content. The problems of not verifying each post will be discussed in chapter 7. The collection of the extreme dataset was performed in collaboration with Fernandez.

5.2.3 Radical Data

The radical data was retrieved from Gab using the same approach as when collecting the extreme data. The difference was which users' posts to keep. Since the users that were collected as described in the first part of the previous subsection all commented on the known extreme user's post, it was concluded that most of them would not be defined as neutral users. Two hundred and five of the users that were not banned from Twitter were randomly chosen to be part of the radical dataset. News users and other users that obviously were not radical were not added. When only keeping the original, non-duplicate gabs with less than 300 characters, it remained about 67,000 posts. Likely, there are users in the radical dataset that could be classified as extreme and visa versa. However, it was decided to try this approach in this experiment and leave the improvement of datasets to future work. Differences between the two datasets can be found in section 6.1 where the results of the text analysis are presented. Drawbacks of this approach will be discussed in chapter 7.

5.2.4 Data for Prediction

Three smaller datasets were collected for prediction after the ML models were trained. The data used as neutral was found on Kaggle and is called "Tweets

Dataset"[5] and contains posts of the 20 most popular Twitter users in 2017. For extreme prediction data, the posts of users banned from Twitter that have commented on the mentioned post in subsection 5.2.2 but not included in the extreme dataset were scraped. The last one containing radical users was gathered by retrieving not Twitter banned users that had commented on the *"Help find me find my frens"* post and that were not used in the radical dataset. The extreme and radical datasets consisted of posts by 167 users, where 43 were defined as extreme and 124 as radical. Eight thousand posts, equally distributed between the three categories, were used as the prediction dataset.

5.3 Text Analysis

Before using NLP and ML techniques, the retrieved data was analyzed. It was considered interesting to see if it could be found differences in post length, most used words, hashtags and mentions between the assumed extremists and radicals and the neutral users. Lastly, a brief analysis of the use of URLs was performed.

5.3.1 Length

Lara-Cabrera et al. [2019] had a hypothesis that radicalized Islamists tended to write shorter posts than other users. They stated that introverted people tend to post shorter text than others, and they expected the radicalized users to be more often introverted than the general user. Surprisingly, their results showed the opposite; The radicalized users posted longer posts than the others. Ahmad et al. [2019] got the same results as Lara-Cabrera et al., extremist post longer post (10.47 words) than non-extremists (8.92 words). Therefore, the average and median length for the extreme, radical and neutral datasets were checked to see if it was possible to see the same pattern for right-wing extremists. The lengths were calculated on the posts after URLs and emojis were removed.

5.3.2 Frequent Words

Before counting the frequencies of words in each corpus, the text was preprocessed as described in section 5.4. Preprocessing was performed because it was important to return typical or commonly used words in each corpus, not frequently used English words. Contrary to the rest of the subsections in section 5.3 the text had to be preprocessed. However, since most of the text analysis tasks were performed on the raw text, it was decided to introduce all of the analysis processes before how preprocessing was performed. The most frequent words used

[5] The dataset is available here: `https://www.kaggle.com/datasets/mmmarchetti/tweets-dataset?resource=download` (Accessed: 05.05.22).

for each dataset were found by iterating through each corpus. The results is displayed in Table 6.2a.

5.3.3 Hashtags and Mentions

It was searched through the three datasets to map the use of hashtags and mentions to find the most used ones and compare the results for each dataset. The analysis investigated whether hashtags and mentions could be promising identifications for recognizing people vulnerable to radicalization on social media platforms. Hashtags are used to emphasize, for example, a case or feeling, but they are more used to get the post to appear in desirable groups or searches [Berger and Bill, 2013]. Agarwal and Sureka [2015] states that the hashtags deliver significant indicators of the theme of a post. Subsection 6.1.3 discussed the used hashtags and mentions in the neutral, radical and extreme datasets. Hashtags and mentions were not used as features in the ML models but are considered relevant to include in future work.

5.3.4 URLs

URLs are not considered features in the experiments. However, it was decided to investigate the frequency of posts in the three datasets that contained URLs.

5.4 Preprocessing data

After the text analysis experiment was finished (except subsection 5.3.2), the datasets were ready for preprocessing. NLTK and Num2Words, presented in section 3.6, were used. The approach consisted of the six following steps. The process is described with an example using the post *"Today I had a nice day. I found 7 fun videos at https://www.google.com."*.

1. Change all letters to lowercase.
   "today i had a nice day. i found 7 fun videos at https://www.google.com.".

2. Remove URLs and emojis.
   "today i had a nice day. i found 7 fun videos at.".

3. Convert digits to words.
   "today i had a nice day. i found seven fun videos at."

4. Remove punctuation, symbols and stop words.
   "today nice day found seven fun videos"

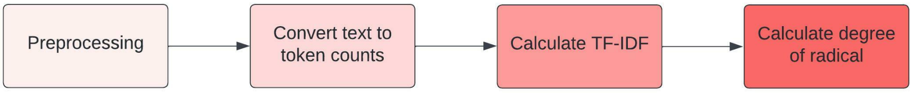


Figure 5.2: The four steps of the process of calculating the TF-IDF score and using it to calculate how radical posts are.

5. Tokenize each document.
   [today, nice, day, found, seven, fun, videos]

6. Perform lemmatlization on each token.
   [today, nice, day, find, seven, fun, video]

The preprocessing transforms a post from a string (sentence) to a sequence (list) of terms that are easier to analyze and compare with others.

5.4.1 TF-IDF

The Sklearn library was used to calculate the TF-IDF scores. It utilized CountVectorizer and TfidfTransformer to create a vocabulary and then calculate the TF-IDF scores. It was decided to try different approaches when assigning the weights of each word. The first one was the standard variant, where it is expected that words with a low appearance in the corpus deliver the most meaning to a document, while the second approach valued the more frequent words the most. The TF-IDF scores were used to calculate how radical a post is. Figure 5.2 shows the process that was implemented to calculate the percentage of how radical a post is using TF-IDF. This process applies both when valuing the rare words and when valuing the common words. CountVectorizer is used in step two, and TfidfTransformer is used in step three. The 50 words given the highest IDF scores were collected and used to create a dictionary.

5.4.2 Word2Vec

The Word2Vec experiment aimed to visualize similarities between words in the extreme dataset. It was desired both to plot the 50 most frequent words in a plot and to get a list of the ten most similar words to each of the top 50 words according to the best model. Multiple configurations were tested, but two of them were chosen to include in the thesis since they gave the most meaningful results.

Table 5.1: The three layer model.

Layer (type)	Output Shape	Param #
Input layer (Dense)	(None, 32)	13777088
Dropout (Dropout)	(None, 32)	0
Hidden layer (Dense)	(None, 16)	528
Dropout (Dropout)	(None, 16)	0
Output Layer (Dense)	(None, 3)	51

- **Model One:** The first word2vec model was built with a *min_count* of ten and a *vector_size* of 50. The model was trained with 10,000 epochs.
- **Model Two:** For the second model, it was tested if a *min_count* of seven and *vector_size* of 50 trained for 200 epochs could deliver good results.

5.5 ANN Multi-Label Classifier

This section provides information on how ANNs were built, trained and tested. The parameter tuning performed in the developing process is described in this section. It was developed different versions of ANNs for multi-label classification to test if they could deliver promising results for predicting if social media posts are *neutral*, *radical* or *extreme*. As for TF-IDF, CountVectorizer was used to create a matrix representation according to the BoW technique. Keras was used to build the model with multiple layers.

5.5.1 The Used Neural Networks

Different configurations of layers and output shapes were tried, but Table 5.1 and Table 5.2 were the architectures that delivered the most promising results and therefore included in the thesis. Table 5.1 only consist of one hidden layer, while Table 5.2 added one extra. The choice of only including one hidden layer was to reduce the risk of overfitting. Because of the risk of overfitting, a dropout layer was added after all the layers except the output layer. The rate for dropouts did vary for the models but was for all models set to around 0.5. The dropout layer changes the layer's input to zero instead of the original value. This is done at the rate defined.

5.5.2 Parameter Tuning

Multiple configurations for models were tested and trained, but it was chosen to include the four trained models below because they delivered the most promising

Table 5.2: The five layer model.

Layer (type)	Output Shape	Param #
Input layer (Dense)	(None, 64)	13777088
Dropout (Dropout)	(None, 64)	0
Hidden layer (Dense)	(None, 32)	2080
Dropout (Dropout)	(None, 32)	0
Hidden layer (Dense)	(None, 16)	528
Dropout (Dropout)	(None, 16)	0
Output Layer (Dense)	(None, 3)	51

results, which will be presented in section 6.3. The reason for the number of epochs in the different models is explained in subsection 5.5.5. It was considered to try *Kullback Leibler Divergence* as an optimizer, but after testing, it was decided that it should be dropped.

Model One

Architecture: Table 5.1
Optimizer: SGD
Batch size: 125
Loss: Sparse Categorical Crossentropy
Epochs: 100

Model Two

Architecture: Table 5.2
Optimizer: SGD
Batch size: 64
Loss: Sparse Categorical Crossentropy
Epochs: 44

Model Three

Architecture: Table 5.2
Optimizer: SGD
Batch size: 248
Loss: Sparse Categorical Crossentropy
Epochs: 72

Model Four

Architecture: Table 5.1
Optimizer: Adam
Batch size: 125
Loss: Sparse Categorical Crossentropy
Epochs: 150

5.5.3 Splitting of Data

The extreme, radical and neutral datasets were merged, creating a dataset for creating and training the model. Since the ML approach is supervised, each post needed to have the corresponding value representing its belonging category. Further, the model required a test set. The division of the dataset was performed using the code in Listing 5.1. The *train_test_split* method divides the data randomly by shuffling the data before dividing the dataset. The shuffling is controlled by the parameter called *random_state*. 75% of the data was used to train the model and the remaining 25% for training.

```
sentences_train, sentences_test, y_train, y_test = train_test_split(
    sentences_x, sentences_y, test_size=0.25, random_state=1000)
```

Listing 5.1: Splitting dataset to train and test.

5.5.4 Building the Model

The models were built with layers as shown in subsection 5.5.1 and with the specifications described in subsection 5.5.2. Further, the models were compiled using the code in Listing 5.2. The three defined parameters for the compile method specify the loss function, the optimizer and the evaluation method to use, respectively. The different applied choices for the parameters are presented in section 3.4.3. In all the models used in the experiment, the schedule in Listing 5.3 was implemented to define the learning rate.

```
import keras

#Compile the model
model.compile(loss, optimizer, metrics)
```

Listing 5.2: Compiling the model.

```
import keras

#Implement a learning rate schedule
```

```
lr_schedule = keras.optimizers.schedules.ExponentialDecay(
    initial_learning_rate, decay_steps, decay_rate)
```

Listing 5.3: Learning Rate Schedule.

5.5.5 Training the Models

When the models were built and compiled, they were ready to be trained. Listing 5.4 show the training, or fitting, method. The training and validation data are the data that were split in Listing 5.1 but transformed for the Keras model to be able to understand it. The three first models in subsection 5.5.2 were configured to be trained for 100 epochs and the last one for 150 epochs. To prevent the model from training after it has reached its potential, it was implemented a function, which is shown in Listing 5.5, to make the model stop training when the validation loss increases for more than ten epochs.

```
import numpy as np

# Train the model
history = model.fit(training_data, epochs, verbose, validation_data,
     batch_size, callbacks)
```

Listing 5.4: Fitting the model.

The variable *verbose* lets the programmer choose to show the progress as it is training. *Batch_size* is presented in subsection 3.4.3 describing what it is and which value chosen for each model.

```
import keras

#Stop the training when the validation loss is increasing
es_callback = keras.callbacks.EarlyStopping(monitor=’val_loss’,
    patience=10)
```

Listing 5.5: Function to stop the training when the validation loss is increasing.

5.5.6 Prediction

The models were then evaluated using the prediction dataset to calculate evaluation metrics and analyze specific users to see the distribution of posts. The predictions are visualized in confusion matrices. The diagonal of the matrix should have the most predictions since the values on the diagonal represents the correct predictions.

Evaluation Metrics

Accuracy, Precision, Recall and F1 scores were calculated for each class in the models using the prediction dataset. The evaluation metrics should be used to minimize the number of neutral users predicted as radical or extreme but also to maximize the overall accuracy.

Pie Diagrams for Users

Pie diagrams were created to visualize how a user's social media posts were classified. The visualization aimed to see if the approach could be a promising method to recognize if the user started acquiring radical beliefs. Possibly it could be calculated for a user in different periods of time and compared to see if the distribution changed over time. The latter task was not performed in the experiment but is considered an interesting task for future work.

Chapter 6

Results

This chapter presents the results of the experiments. First, the findings for the text analysis are presented, followed by the word embedding results. Lastly, the findings from the ML experiment are presented. This chapter only presents the results, while they will be discussed and evaluated in chapter 7.

6.1 Text Analysis

The goal of the text analysis was to get an overview of typical characteristics in the structure and language of posts for the three groups of people. First, the length of the posts was calculated, followed by the most commonly used words in the datasets. Lastly, an analysis of the usage of mentions, hashtags and URLs is presented.

6.1.1 Length

In Table 6.1 the average and median word and character length for the three datasets are displayed. Only the gabs with less than 300 characters were included in the analysis for the extreme and radical datasets. The extreme posts were significantly longer for both the word and the character length than the radical and neutral posts. A reasonable guess would be that the lengths for the radical dataset would be in the middle of the extreme and neutral datasets. However, it turns out that the radical users in the collected dataset write fewer words than both the other groups. The variations between the average and median length for both the extreme and the radical posts were larger than for the neutral posts.

	Word length	*Character length*
Extreme		
Average	16.3	100.4
Median	13	84
Radical		
Average	11.1	72.6
Median	8	52
Neutral		
Average	12.6	71.0
Median	11	66

Table 6.1: The average and median word and character lengths for the three datasets.

6.1.2 Frequently Used Words

The most used terms are presented in Table 6.2 with the percentage of documents each term appears in. The most frequent used words in the neutral dataset could be considered ”normal”, not delivering any meaningful information to the reader. Compared to these neutral words, the radical and the extreme most frequent words deliver more information to a sentence. The majority of the words in Table 6.2a can be considered to be helpful words to consider for further analysis. Using the words from Table 6.2c would probably not help classify users. As

Term	**Present in posts**
Trump	7.4 %
Democrat	5.2 %
Like	5.1 %
Maga	5.0 %
Kek	5.0 %
People	4.8 %
White	4.5 %
Trudeau	3.7 %
Meme	3.6 %
Get	4.0 %

(a) The extreme dataset.

Term	**Present in posts**
Like	4.8 %
Im	3.4 %
People	3.4 %
Get	3.4 %
One	3.3 %
Gab	3.0 %
Dont	3.0 %
White	2.5 %
Time	2.5 %
Biden	2.2 %

(b) The radical dataset.

Term	**Present in posts**
Im	10.0 %
Day	6.1 %
Good	5.1 %
Get	4.9 %
Go	4.6 %
Like	4.6 %
Work	4.3 %
Dont	4.1 %
Today	4.1 %
Cant	3.9 %

(c) The neutral dataset.

Table 6.2: The most frequent used words in the three datasets.

Popular Hashtags					
Extreme	%	Radical	%	Neutral	%
#maga	4.84	#thotwars	0.29	#followfriday	0.17
#kek	4.67	#impeachbiden	0.27	#fb	0.89
#trump	4.23	#traitorjoe	0.22	#asot400	0.05
#democrats	4.10	#afpac	0.20	#ff	0.04
#cdnpoli	3.51	#wakeupamerica	0.12	#delongeday	0.01

Table 6.3: The five most popular hashtags in the three datasets and the percentage of posts they appear in.

suspected, the most frequent radical terms in Table 6.2b consist of a mix of "normal" and more domain-specific words.

6.1.3 Mentions and Hashtags

45.3% of the posts in the neutral dataset contained at least one mention, and 2.1% of them had at least one hashtag. The extreme dataset's corresponding values were 7.6% and 39.6%, respectively. 3.9% of the radical dataset contained at least one mention, and 5.0% had at least one hashtag. Table 6.3 shows the most frequently used hashtags in the three datasets. Word-clouds for each of the three datasets are displayed in Figure 6.1. Except for the most mentioned user in the extreme dataset, the frequency of the most used mentions appears similar for all three groups. Due to privacy, neither tables nor word-clouds with the most frequent mentions are included.

The usage of hashtags was similar by neutral and radical users, with only a little more frequent use of hashtags by the radical than the neutral users. Compared to these two groups, the users in the extreme dataset uses it to a much larger extent. *Maga* ("Make America Great Again"), *Trump*, *Democrats* and *Cdnpoli*[1] are all hashtags related to politics, whereas *kek* is used as "lol"[2] by people playing World of Warcraft. Since four out of five of the most used hashtags in the extreme dataset are related to politics, it suggests that the extreme users express a lot of political beliefs. Except from the hashtag *thotwars* all radical hashtags in Table 6.3 are related to politics as for the extreme dataset. *Thot* is used as slang for a woman considered to be a "hore" or "slut"[3]. Knowing the definition of *thot*,

[1] #cdnpoli is a well known hashtag to use when writing about the Canadian politics.

[2] Lol stands for "laugh out loud".

[3] *Thot*: *"a woman considered to be sexually provocative or promiscuous; a slut or whore."* according to dictionary.com (`https://www.dictionary.com/browse/thot` (Accessed: 24.05.22)

Term	IDF Weight
Trump	8.564
Maga	8.038
Like	8.026
Democrats	7.995
Kek	7.983
People	7.962
White	7.902
Covid	7.843
Trudau	7.772
Cdnpoli	7.713

Table 6.4: IDF scores frequent used words in the extreme dataset.

it is expected that the hashtag *thotwars* is used by incels that want to combat women. None of the top five neutral hashtags is related to politics.

6.1.4 URLs

The number of posts in the extreme dataset that contained an URL was 19.5%, while 16.5% of the posts in the radical dataset included an URL. The neutral dataset had a lower percentage of posts containing URLs than the other two groups. The number of neutral posts that included an URL was 3.9%. When said that a post includes a URL, it means that it has at least one URL, but it can contain more than one URL.

6.2 Text Embedding

This subsection includes the results of the word embedding techniques applied on the extreme datasets. First, the TF-IDF findings are presented before results from the training of the Word2Vec models are explained and shown.

6.2.1 TF-IDF

It is suspected that posts in other datasets than the extreme dataset used to calculate the IDF scores would get a lower radical degree due to the rapid change of focus of political subjects. For example, it is expected that after the war started in Ukraine, the war-related content on social media would explode since many of the extreme Gab users probably would post radical beliefs about the war. Despite this, the TF-IDF scores might suggest that they are neutral because the

Category	**Most Extreme**	**Not Extreme**
Extreme	*gab*	*The Taliban are banning Tiktok! This is amazing*
Radical	*MAGA*	*Russian Ministry of defense claims that Ukrainian airbases and air defense destroyed*
Neutral	*Twitter;;;;;;*	*Quit being a quitter.*

Table 6.5: The least extreme and most extreme posts for each category in the predicting dataset.

words used to determine the IDF scores were different from the words used when posting about Putin, Russians and possibly nuclear weapons.

Table 6.4 shows the transformed IDF scores for the ten most frequent words in the extreme dataset. The standard TF-IDF algorithm would value the rare words. The new IDF scores were transformed so that the most frequently used words had the highest value, and the rare the word, the lower the score. The lowest score given was 0.405. Only the 500 words with the highest IDF scores were used to calculate the TF-IDF scores for posts. These 500 words created a dictionary which is included in section E in the appendix. Since the values calculated using the ordinary TF-IDF equation not returned results considered accurate, the values for that approach are not included in the thesis.

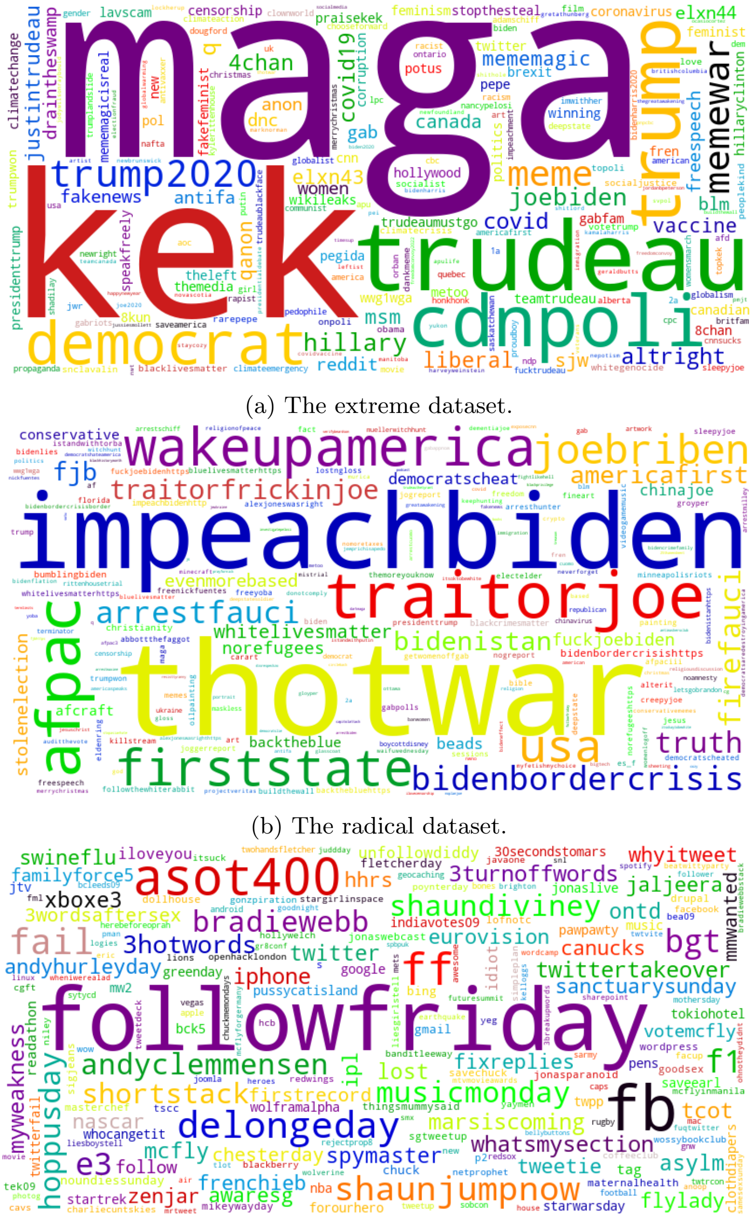
An array was returned for each post analyzed using the IDF scores. A value was calculated for posts representing the degree of how radical the post is. The approach was calculating the average value, but that favored the short posts. Future work should explore alternatives so the length does not affect the result as much as the current solution.

Table 6.5 displays the post that received the highest score of extreme and one post that received a value of zero. The table shows that short posts that include words that are in the dictionary in section E in the appendix will be given a high value. The posts in the not extreme column in the table are not very long, probably since the longer the posts, the more chance that it includes at least one word in the dictionary. Long posts will not get a high value but, in most cases, nor be given a value of zero. Of the radical posts, most of the posts with a value of zeros were URLs with no additional text.

6.2.2 Word2Vec

Figure 6.2 shows a plot for the two models, presented in subsection 5.4.2, trained to represent each word as a vector to see similarities between words. The plot includes the 50 most frequent words in the dictionary in section E in the appendix. In Figure 6.2a and Figure 6.2b the words *Justin, Trudeau, Canada, Liberals, Cdnpoli* and *Elxn* appears close to each other suggesting that they often appears in similar situations. All of them are related to the Canadian election. In Table 9 in the appendix, the top 50 used terms in the extreme dataset are presented with the top ten most similar words according to the first Word2Vec model.

Figure 6.1: Most used words hashtags presented in a word-cloud for each dataset.

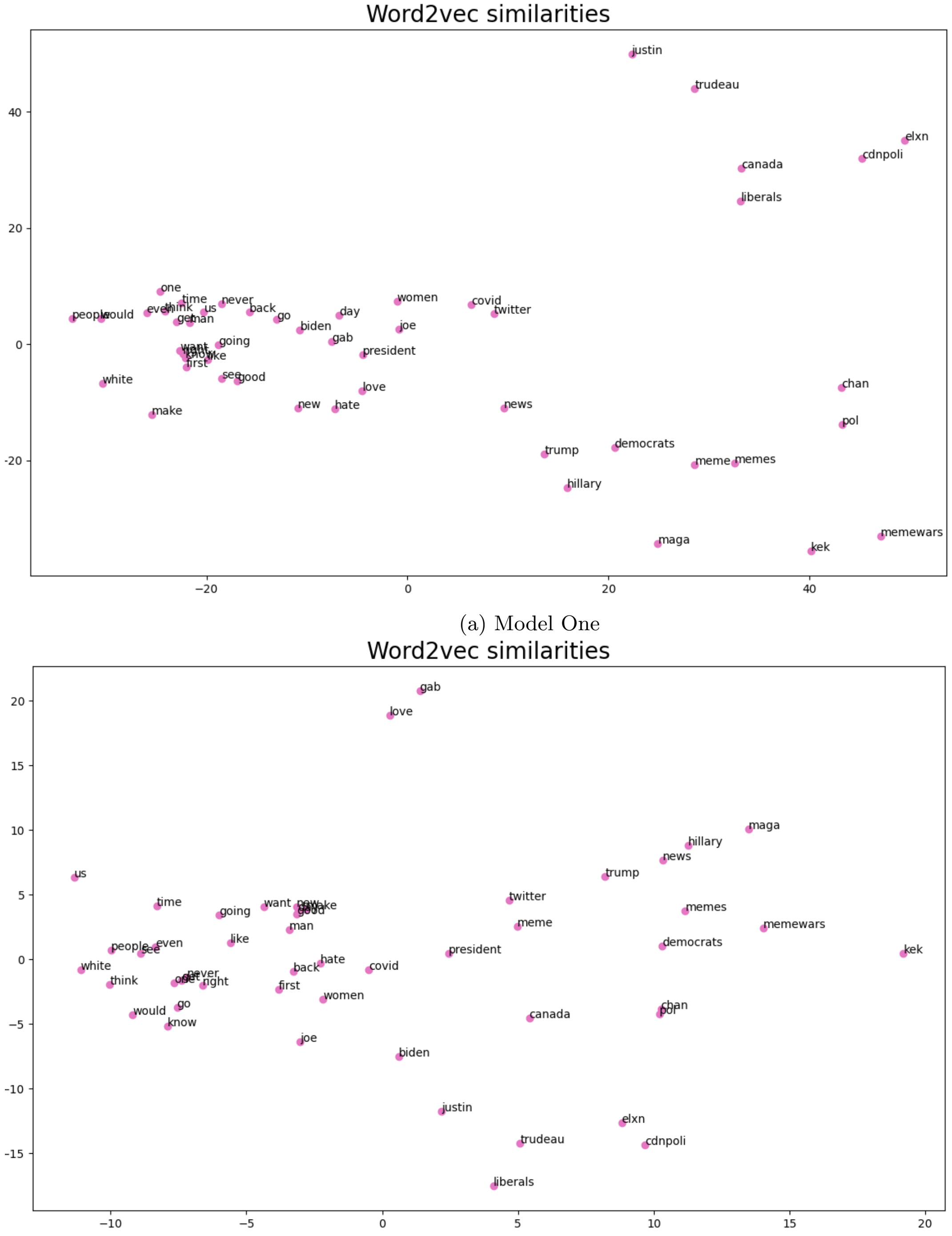


(a) The extreme dataset.

(b) The radical dataset.

(c) The neutral dataset.

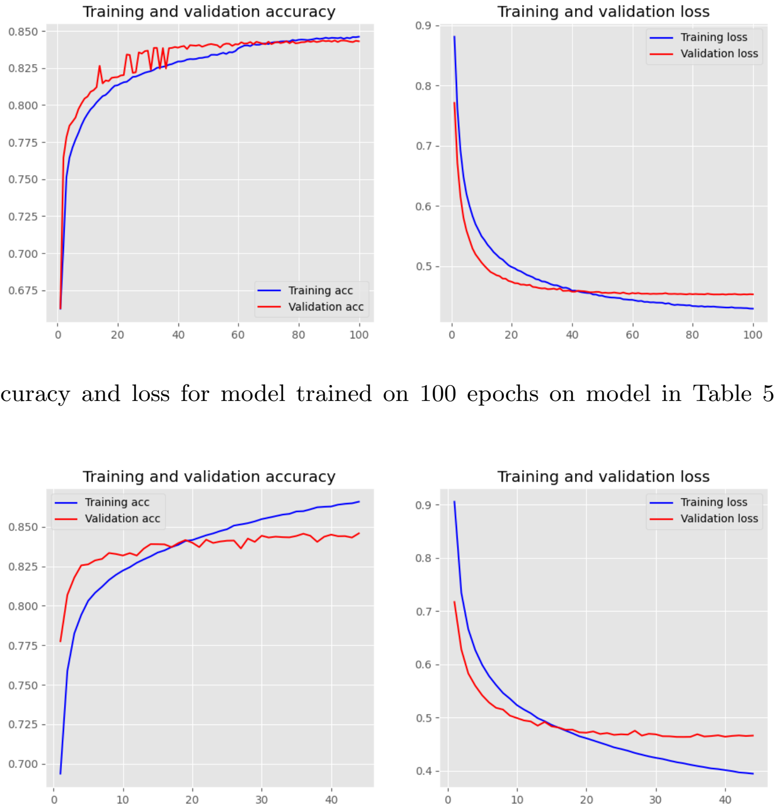
Figure 6.2: Plot of 50 most frequent words in the extreme dictionary plotted using to different word2vec models.



(a) Model One

(b) Model Two

Figure 6.3: Loss and accuracy for the trained models.



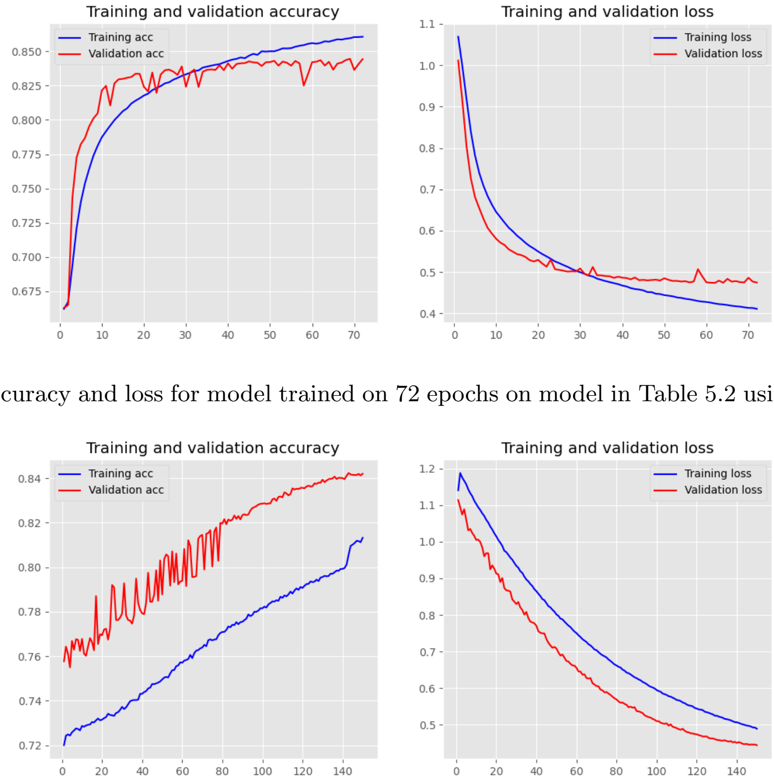
(a) Accuracy and loss for model trained on 100 epochs on model in Table 5.1 using SGD.

(b) Accuracy and loss for model trained on 44 epochs on model in Table 5.2 using SGD.

6.3 Deep Learning

This section presents plots of accuracy and loss for each training of the four models in section 5.5. For each trained model, TP, FP, TN and FN are displayed in a confusion matrix in Figure 6.5. Thereafter, the models are evaluated using matrices before they are used to visualize users' posts in pie diagrams.

Figure 6.4: Loss and accuracy for the trained models..



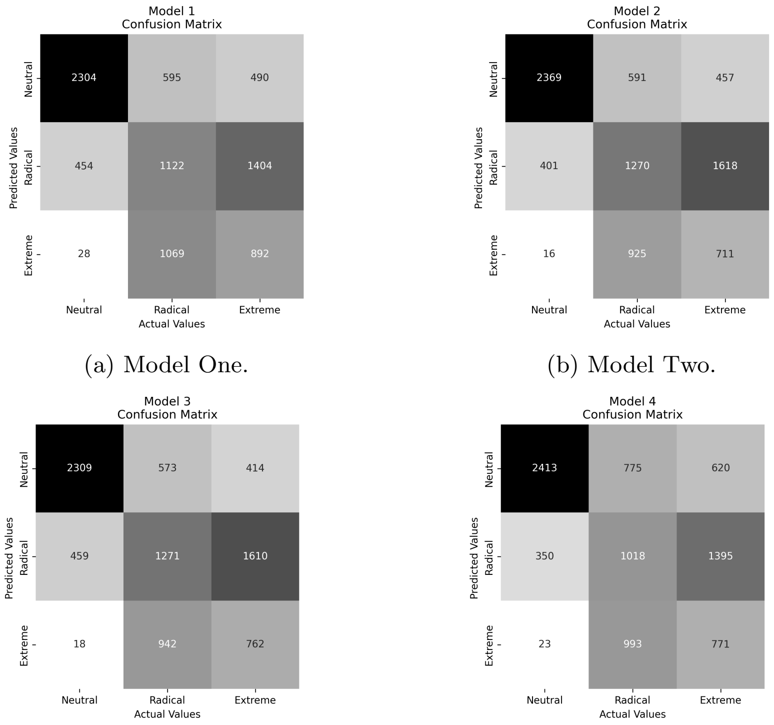
(a) Accuracy and loss for model trained on 72 epochs on model in Table 5.2 using SGD.

(b) Accuracy and loss for model trained on 150 epochs on model in Table 5.1 using Adam.

6.3.1 Plots

The plots in Figure 6.3 and Figure 6.4 shows the training process for each of the models. Each training session is visualized with two plots, one for the training and validation accuracy and one for the training and validation loss. Model 2 in Figure 6.3b returned the best accuracy for both training and validation.

Figure 6.5: Confusion matrices for the trained models.



Model 1
Confusion Matrix

Predicted Values \ Actual Values	Neutral	Radical	Extreme
Neutral	2304	595	490
Radical	454	1122	1404
Extreme	28	1069	892

(a) Model One.

Model 2
Confusion Matrix

Predicted Values \ Actual Values	Neutral	Radical	Extreme
Neutral	2369	591	457
Radical	401	1270	1618
Extreme	16	925	711

(b) Model Two.

Model 3
Confusion Matrix

Predicted Values \ Actual Values	Neutral	Radical	Extreme
Neutral	2309	573	414
Radical	459	1271	1610
Extreme	18	942	762

(c) Model Three.

Model 4
Confusion Matrix

Predicted Values \ Actual Values	Neutral	Radical	Extreme
Neutral	2413	775	620
Radical	350	1018	1395
Extreme	23	993	771

(d) Model 4.

6.3.2 Confusion Matrices

The four confusion matrices in Figure 6.5 are based on the predictions made by each model on the prediction dataset. The matrices give a simple visualization of the performance of each model and which types of predictions it performs best. When it is essential to reduce the number of neutral users accused of being radical or extreme, it is desired that the second and third rows in the first column are as close to zero as possible. Additionally, the diagonal representing the three cases where the predictions are correct should be the darkest in the figure.

6.3.3 Evaluation Metrics

The evaluation metrics for the four models using the dataset gathered for prediction and evaluation, see subsection 5.2.4, are shown in Table 6.6. *Accuracy*, *Precision*, *Recall* and *F1-Score* were calculated for all three possible categories. Models 2 and 3 received the best accuracy for prediction (52.0%) of the four models. The models that received the best average value for precision, recall and F1 score are marked in grey. Model 1 turned out to be the worst. The recall for the neutral class is an important metric to inspect when it is desired to reduce the number of neutral users classified as extreme or radical. Model 4 delivered the best result of recall (86.6%) for the neutral users.

6.3.4 Predictions of Users

Posts of two users from each of the categories in the predicting dataset were visualized in pie diagrams, using model 4, in Figure 6.6. Model 4 was chosen since it delivered the best results when focusing on classifying the neutral posts correctly. It was chosen users that had posted several hundred posts, so the basis for the comparison should be similar. The two neutral users in Figure 6.6a and Figure 6.6b are Cristiano Ronaldo and Katy Perry, respectively. Both of them have public positions and are therefore mentioned by name. The radical and extreme users in Figure 6.6 are not mentioned by name due to privacy. They are not considered public people like Ronaldo and Perry, and none of them has a know position.

Donald Trump was not one of the users in the prediction dataset, but since he is banned from Twitter, it was chosen to include him to compare him to Barack

	N	**R**	**E**	**Ev**
Total Accuracy: 51.7%				
P in %	68.0	37.7	44.8	50.2
R in %	82.7	40.3	32.0	51.7
F1 in %	74.6	43.6	33.7	50.6

(a) Model 1

	N	**R**	**E**	**Av**
Total Accuracy: 52.0%				
P in %	69.3	38.6	43.0	50.3
R in %	85.0	45.6	25.5	52.0
F1 in %	76.4	54.9	24.8	52.0

(b) Model 2

	N	**R**	**E**	**Av**
Total Accuracy: 52.0%				
P in %	70.1	38.1	44.3	50.8
R in %	82.9	45.6	27.4	52.0
F1 in %	76.0	53.0	27.0	52.0

(c) Model 3

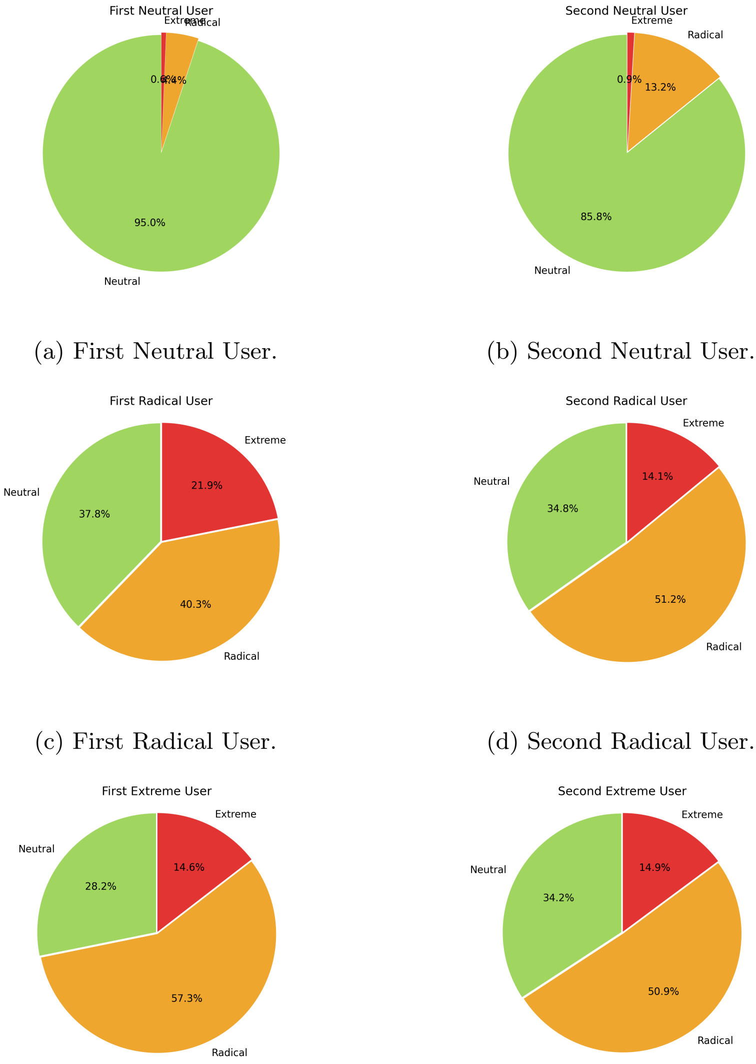
	N	**R**	**E**	**Av**
Total Accuracy: 50.3%				
P in %	63.4	36.8	43.1	47.8
R in %	86.6	36.5	27.7	59.3
F1 in %	73.2	41.6	30.0	48.3

(d) Model 4

Table 6.6: Evaluation metrics for the four models in percentage. *N=Neutral, R=Radical, E=Extreme, Av=Average, P=Precision, R=Recall, F1=F1-Score*

Obama, who is one of the users in the neutral dataset. The comparison of Trump and Obama for all four models is visualized in Figure 6.7. Donald Trump and Barack Obama have public positions and are therefore considered reasonable to mention them by name. It is crucial to state that the predictions in Figure 6.7 of Donald Trump and Barack Obama are solely based on the prediction models, not the author's opinions about either one of them and should not be used as facts.

Figure 6.6: Pie diagrams for the six chosen users for model 4.



(a) First Neutral User.

(b) Second Neutral User.

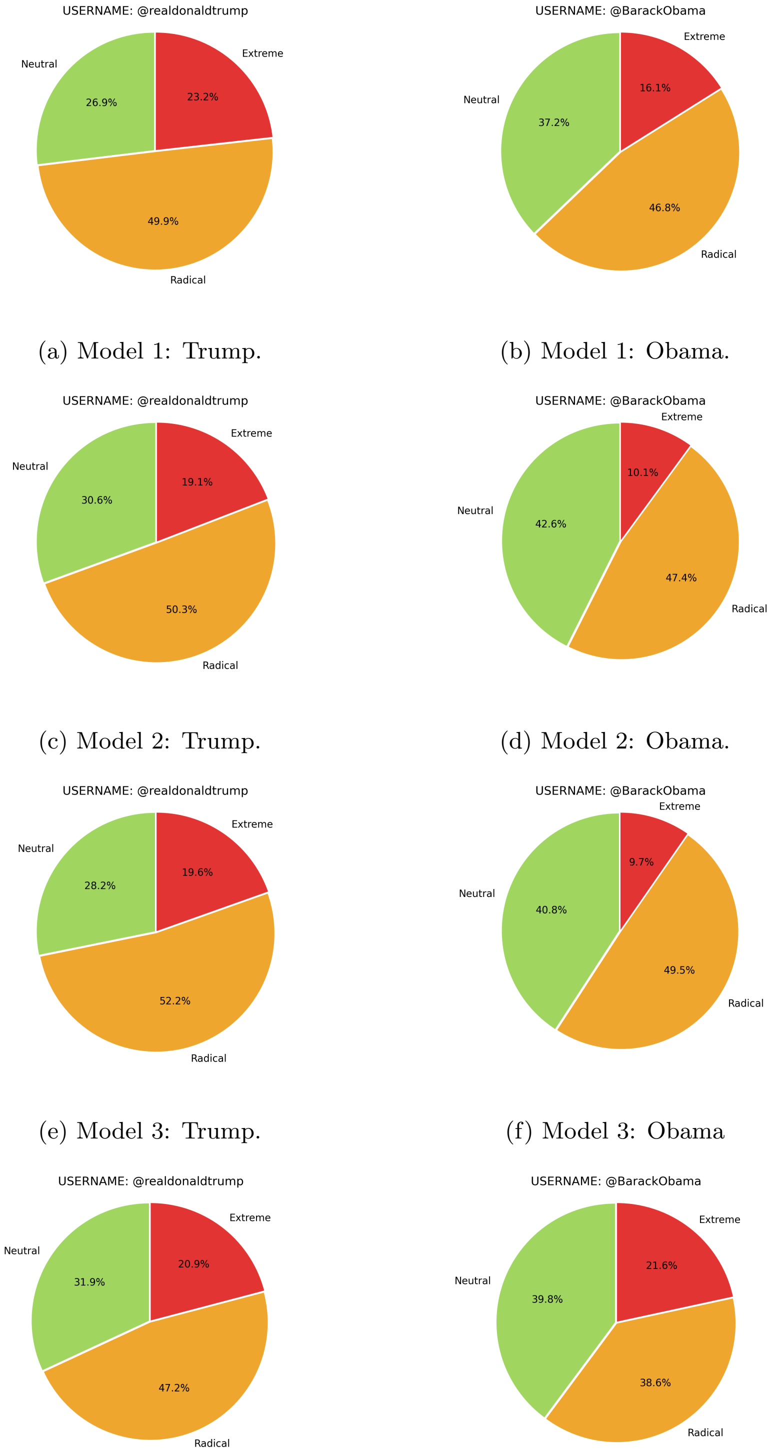
(c) First Radical User.

(d) Second Radical User.

(e) First Extreme User.

(f) Second Extreme User.

Figure 6.7: Pie Diagrams for Donald Trump and Barack Obama for the four models.



(a) Model 1: Trump.

(b) Model 1: Obama.

(c) Model 2: Trump.

(d) Model 2: Obama.

(e) Model 3: Trump.

(f) Model 3: Obama

(g) Model 4: Trump.

(h) Model 4: Obama.

Chapter 7

Discussion and Evaluation

In this chapter, the findings in chapter 6 are discussed and evaluated. First section 7.1 and section 7.2 answers the goals and RQs based on the experiments. Then findings not answering the goals and RQs are discussed in section 7.3. Lastly, section 7.4 discusses the approaches used in the experiments.

7.1 First Goal

Goal 1: *Investigate differences between right-wing extremists' and other users' language used on social media based on collected datasets.*

The experiments in section 5.2 and section 5.3 were performed to fulfill the first goal. During the work of the master's thesis, three datasets were used for analysis, and it was found considerable differences in the use of language. It can be concluded that the goal was fulfilled given the chosen datasets and the investigated features. However, as will be discussed in subsection 7.1.1, it can not be concluded that there are universal differences in language use. All findings in the experiments are true for the given datasets, not generally for right-wing extremists and radicals. Additionally, the definition of extremism is crucial since the extreme data may by others be considered radical, and the radical dataset could be considered extreme. The rest of this section and section 7.2 will be discussing the findings based on the given conditions, meaning that it should not be considered to be true for all situations.

7.1.1 Research Question 1

RQ 1: *Can appropriate datasets for social media content for extreme and radical right-wing content and other users be found publicly available or collected* *from social media platforms?*

In total, four datasets were collected. Three were used to train the model, and the latter was used to predict and evaluate the trained models. The fourth dataset consisted of an open-source Twitter dataset and extreme and radical data collected from Gab. The radical and extreme datasets used for training were also collected from Gab using the gabs of users commenting on a post by a known extremist and separated based on who was banned from Twitter and who was not.

The conclusion of RQ 1 is that finding relevant existing publicly available online datasets of extreme and radical right-wing content was not possible on either Twitter or Gab. Therefore, it was chosen to scrape data from Gab since it was considered difficult to find extremists on Twitter since they ban extreme users. When Gab was chosen as a social media platform for extreme content, it was desirable to get a neutral dataset from the same platform. Unfortunately, it was considered problematic and decided that the best solution would be to use an existing neutral Twitter dataset. The metadata is different for Gab than Twitter, but not an essential difference since the textual posts are the only thing analyzed in this project.

As mentioned in subsection 2.3.5, the character limit for gabs had increased after the decision to use Gab was made. Hence the difference in allowed characters increased from 20 to 2,720 characters. Since the experiment used two social media platforms, the posts needed to have a similar structure and limitations to be compared. To minimize the consequences of the character limit change, it was decided to exclude all gabs that contained more characters than 300, which was the previous character limit for gabs. It is challenging to know the consequences of removing an amount of the collected data. However, for the case of this experiment, it is considered not to make the results unusable but that they could be more accurate if all posts were used. The removal of longer gabs did not directly affect RQ 1, but the text analysis that answered RQ 2 may be severely affected. A solution to the problem could be to create platform-specific analysis and ML models. Gathering neutral data from Gab is expected to be more difficult than on Twitter. However, it should be investigated if a large enough dataset of neutral posts could be collected from Gab to see if the models' accuracy was improved.

As of May 2022, it is impossible to see group members on Gab and usernames of people liking posts. If that was possible, it could be easier to understand a user's beliefs and recognize if a user can be considered neutral.

After finishing the experiments, Gab is considered a platform suitable for recognizing people in the radicalization process since the average of the investigated

Gab content seemed more right-wing oriented than social media platforms like Twitter and Facebook. Different approaches should be investigated to define someone as radical or extreme if reusing Gab. Given the results, especially the text analysis in section 6.1, the content of the extreme dataset seemed more right-wing extreme than the radical and the neutral dataset, suggesting that the annotating process could work. However, since all data posted by a user that was banned from Twitter was considered to be extreme, the extreme dataset likely contains posts that should be classified as radical or neutral. To manually go through every post to check if it was classified correctly would be time-consuming and challenging without more domain knowledge.

The retrieved posts from Gab were only collected from a scope of possible extremists and radicals since they all followed the same user, suggesting that all of them belong to the same or similar communities and have similar beliefs. It is not analyzed where the users are located, so it is not possible to say if they represent people holding right-wing beliefs common worldwide or if it is more place-specific, like if the majority of the collected users are stationed in the US. Skleparis and Knudsen [2020] presented that the variations in right-wing content differ from the UK to Greece, which makes it likely that there exist variations between countries and places in general. Right-wing extremism includes different groups of members, like white supremacists, incels and neo-Nazis, which targets different victims as presented in section 2.2. The experiments do not investigate which type of extremism is present in the dataset because all types of right-wing extremism are considered the same in the experiments. Hence, the results could be accurate for white supremacists but not neo-Nazis.

The posts in the datasets are from different periods. The radical and extreme dataset, both for training, testing and predicting, collected in this work contains posts from 2005 until the start of 2022. The neutral dataset used when building and training the models was gathered in April, May, and June 2009. The neutral dataset used for predicting contained tweets from 2011 to 2016. The posts are collected from a period of 17 years, and the time span variations vary significantly. A result may be that the content in the posts in the datasets within a short period contains more similar content than in the dataset that contains posts from different decades. Especially when it comes to right-wing content that contains much political content, the posts in the early 2000s likely discuss other topics than in the 2020s.

The evaluation metrics did not deliver desirable results in Table 6.6 (p.72). However, it is suspected that since the data from Gab collected for the prediction dataset was collected two and a half months later than in the other extreme and radical dataset, the war in Ukraine may be a cause. The war started after the first datasets were gathered. Therefore it is suspected that the posts gathered after contain a more significant percentage of content related to such as war, Putin and nuclear weapons.

The final answer to RQ1 is that it is possible to collect relevant datasets for the three categories when combining novel and existing datasets. However, knowing the conditions considered when collecting the data and metadata of the datasets are vital for interpreting the results correctly.

7.1.2 Research Question 2

When answering RQ 2, the discussion of RQ 1 is essential since all the results from the text analysis rely on the analyzed data. Due to the limitations of the dataset, is it hard to conclude patterns for features on a general basis. However, the text analysis results will be discussed in this subsection for the chosen data.

RQ 2: *Does extreme right-wing users use language features different than other users?*

It was, and still is, considered beneficial to analyze the text before performing any NLP or ML techniques because the results of an analysis most likely will give helpful domain knowledge, improving the following steps like ML. Therefore a significant part of the time used for the experiments went to the analysis. The investigated features were character and word length, most frequently used words, mentions, hashtags, and the frequency of URLs.

The analysis showed that the extreme users' posts consisted of more words and characters than the radical and neutral users' posts. It was not surprising since the same tendency was found by Lara-Cabrera et al. [2019] and Ahmad et al. [2019]. What was surprising was that the number of used words by neutral users was higher than for radical users, while the character lengths only differed by 1.6 characters. It could be suspected that gabs generally would be longer than tweets since the character limit for gabs is higher than for tweets, but then it should have been the case for both the extreme and the radical dataset, not only the extreme dataset. A next step could be to gather more data and investigate if a sign of someone in the radicalization process could be the changes in the length of written posts over time. Could an indicator of users in the process be them starting to write shorter posts than previously and then increasing the posts' lengths when turning to extremists?

The tendency of frequently used words in the tree groups seemed to be represented by a sequential process of using neutral words to using more domain-specific words such as *Gab*, *White* and *Biden*. These words are not normally among the most frequently used words by the population. For the extreme dataset, the only neutral[1] words in Table 6.2a (p.62) are *like* and *get*. The results suggest that analyzing a user's language could predict if a user should be considered in danger of becoming an extremist. Of course, more research needs to be performed to conclude, but the tendency in this experiment is clear.

The mentions analysis showed that the use of mentions was significantly larger in the neutral dataset with almost 50% compared to under 8% for both the radical and the extreme datasets. There is no obvious reason why the use of mentions should differ so much. It could be coincident, time-dependent, or a typical characteristic of normal Twitter users' posts. The variations are so large that it suggests that it is not just a coincidence. Almost 40% of the extreme posts contained at least one hashtag, while 5% of the radical and under 4% of the neutral dataset used hashtags. The most surprising difference between the percentages is the enormous gap between the extreme and radical datasets. Berger and Bill [2013] stated that hashtags are used to emphasize feelings and to make sure the post appears in desirable searches and groups. The desire to spread a specific belief or cause could be a reason for the frequent use of hashtags for the extreme users but does not explain why the radicals use hashtags in significantly fewer percentages of their posts. Could the frequency of use of hashtags be an indicator for determining if someone is approaching the activation point? It would be interesting to investigate if users typically go from little use of hashtags to using them frequently when becoming an extremist.

The extreme users tended to include an URL in their post more often than neutral users and radical users, but the radical users used it almost as often as the extremes. Just about 4% of the neutral posts contained a URL, while the percentage for the extreme and radical users were about 17% and 20%, respectively. Fernandez et al. [2018] suggested using URLs as a feature to determine if a user is in the radicalization process. The simple experiment performed in this project suggests that the usage of URLs could be a good indicator since it seems like the neutral users use significantly fewer URLs than the radical and extreme users. Further, it could be investigated if the content shared in the URLs differed enough to be used as an indicator.

[1] Neutral in the sense that it could be considered a normal word to use frequently by an average person on social media.

7.2 Second Goal

Goal 2: *Use natural language processing and machine learning to recognize people in a radicalization process of turning into an right-wing extremist.*

The second goal aimed to include NLP techniques and ML approaches for calculating the degree of how extreme a post is and use ML to classify posts representing different levels of extreme. Section 5.4 and section 5.5 presents the experiments implemented to fulfill the second goal. The scope of the goal is large, resulting in limited success in fulfilling it. The experiments used NLP and ML to recognize how extreme a post is and categorize it. However, it was not decided a number that represented the activation point or an explanation of what the value given for a post told about how the extreme the post is, except that the larger value, the more extreme. The multi-label classification models classified the posts, not users. Both experiments delivered methods for investigating how extreme posts are, but not methods for predicting the category for users.

Nevertheless, analyzing the posts is considered necessary for further performing similar experiments on a user level. Therefore, it can be concluded that the goal is partly met. RQ 3 and RQ 4 were defined to fulfil the second goal and will be discussed to elaborate on the presented conclusion of the success of the performance to reach the goal. Both research questions are concerned with experiments on the post level, but it was expected that when the posts were classified, it would be a manageable task to define it on a user level. Even though it is not explicitly predicted vulnerable users, the results presented in pie diagrams of different users in subsection 6.3.4 analyze the whole user. However, it is not defined criteria for a user to be neutral, radical or extremist.

7.2.1 Research Question 3

RQ 3: *Can natural language processing techniques be used to calculate a degree of how extreme a post is?*

The main goal of RQ 3 was to use TF-IDF to calculate how extreme a post is. In addition, the NLP technique word2vec was used to see similarities between words used by extremists, but it worked as an analysis of the dataset, not to calculate the degree of how radical a post or user is. The results of the analysis of the frequently used words, presented in subsection 6.1.2, suggested that the frequently used words delivered important information to a post. This assumption is based on the difference in the used words in the neutral dataset, which did not say anything about the discussed topics in the posts, compared to the ones for the extreme dataset, which seemed to tell the topics frequently discussed by the extremists. When implementing the standard version of TF-IDF, the scores of the posts subjectively perceived as extreme were given a low score. Probably since the rarely used words in the dataset were considered to deliver the most information, which is the opposite of what the dictionary text analysis suggested. Therefore it was tried to give the most frequently used words a higher IDF score than the rarely used words, and the results seemed more accurate than the results delivered by the standard version.

A problem occurred when calculating the value of a whole sentence. It was discovered that the short posts were generally given a larger value than longer ones. To calculate the degree of a post, the values for each word in the sentence after preprocessing were summed up and divided by the length of the preprocessed sentence. It should be investigated how the value could be calculated without the length having a significant effect on the score.

The similarities visualized in Figure 6.2 (p.68) using word2vec were only implemented as an analysis of the text. It was interesting that model one managed to place Canada related words in one corner. Model two separated the same words from the rest, but not with as much distance as model one. The rest of the words seemed more randomly placed in both models. The choices of values for the parameters and the dimension for the plot may significantly affect the perceived results of the models. Word2vec is a technique that could be used with ML methods to improve the machine's perception of context.

To conclude, it seems possible to use TF-IDF to calculate the degree of extreme for a post. The approach tested in the experiment should be adjusted further by fine-tuning how the IDF values are calculated and if the TF scores should be converted. Additionally, the corpus used to calculate the IDF values should only contain posts classified and verified as extreme. The results of the performed experiment may be limited due to considering non-extreme words as extreme. The approach to calculating the total value for a post is not optimal, and optimization of the calculation could significantly affect the accuracy of the method.

7.2.2 Research Question 4

RQ 4: *Can artificial neural networks be trained to classify if posts are neutral, radical or extreme?*

Asif et al. [2020] used NB and SVM for multi-label classification and received an accuracy for the classification using SVM of 82%. The models trained in this project were evaluated when using them to predict the classes of the posts in the dataset used for prediction. Model 2 and model 3 received the best accuracy of

52%. The same models gave the best training and validation accuracy, which varied between 84% and 87%, and the training accuracy were higher than validation accuracy for both models. Precision, recall and F1-scores were calculated for each class in the four models.

Recall for the neutral group was considered the most important evaluation metric since the prioritized task of the experiment was to reduce the number of neutral users classified as radicals or extremists. Model 4 got the highest recall for neutral users with a value of 86.6%. Therefore, model 4 was considered the best model in the experiment. It is most important to minimize the wrong classifications for the neutral users since the consequence of accusing someone of being an extremist or in the radicalization process can be severe and, in extreme cases, push them in that direction. Suppose the most important is recognizing all extremists or people in the radicalization, then the precision for these groups should be maximized. However, at the start of developing software to recognize vulnerable people, this should not be the most important. Model 3 would be considered the best if aiming to recognize all radical and extreme posts when considering it most important to not classify them as neutral. When having a model with only a slight change of classifying neutral users wrong, the next step should be to minimize the FN for radical and extreme users.

The confusion matrices in Figure 6.5 show that for all four models, a significant portion of the radical and extreme posts are predicted neutral, which is larger than desired. Since it was decided that all posts by the extreme and radical users were extreme and radical, the predictions may be correct in reality, but the models should still not have classified them as neutral.

The classification in this experiment is performed at the post level but analyzed for users as shown in Figure 6.6 and Figure 6.7 (p.74-75). In Figure 6.6, two random users' distribution of posts from each of the three groups are displayed in pie diagrams. Figure 6.6a and Figure 6.6b shows that both neutral users post less than 1% extreme content and mostly neutral content. The results in these two cases are satisfactory when manually evaluating it subjectively. As described above, the most important evaluation of the model is that they do not predict neutral posts and users to be radical or extreme. In the remanding four pies, the percentage of radical and extreme predicted content increases compared to the neutral users. All four have similar distributions, so it is no clear distinction or pattern between the radical and the extreme users that can be used to recognize people before becoming an extremist. However, since only six users were chosen, the results may be random and not representative for the three groups of users.

Another example used to investigate if the models could be used to predict on a user level can be found in Figure 6.7 (p.75) where the distribution of posts of the former presidents of the US, Trump and Obama, is displayed for the four models. Both former presidents are mentioned by names, and it is described which class they belong to in this experiment based on the defined criteria, not personal opinions about either person. Obama was part of the neutral dataset, while Trump belongs to the extreme class since he is banned from Twitter. Obama's posts have a lower percentage of extreme classifications than Trump's in the three first models, and the percentage of radicals is lower in all cases. Focusing on model 4, since that is the chosen model for the other users, Obama's posts consist of more neutral posts and less radical posts than Trump's, but the differences are minor. The part of the posts that was classified as extreme of Obama's posts was larger than for Trump's posts, but they differed by less than 1%.

It is impossible to conclude why they are so similar, but it is suspected that the right-wing extremists in the extreme corpus discuss a lot of politics, which Obama also does. Obama may post as much political content as those in the extreme and radical dataset. However, Obama's posts are probably considered more acceptable and normal by the average population. The hypothesis that the model assumes political content radical or extreme is strengthen when including the results of Cristiano Ronaldo and Katy Perry in Figure 6.6a and Figure 6.6b (p.74). Ronaldo is a soccer player and Perry an artist making it likely that the percentage of posted political content is significantly lower than for Trump and Obama.

The hyper-parameters should be tuned to improve the evaluation metrics to find optimal values. Before tuning, the architecture of the models should be investigated to look for an optimal number of layers and types of layers. The training and testing data are essential to get an accurate model. As discussed earlier in this chapter, if the annotation process of the posts in the datasets were performed more precisely, the model would probably deliver more accurate results. Another possible reason for less accurate results could be the computer's problem with understanding the context. The ANNs were fed with BoW representation of the text, which does not deliver context information. It could be helpful to use another representation or embedding to take the context into account.

Based on the performed experiments, the short answer to RQ 4 is that it is possible to train ANNs to classify posts. The produced models have the potential for improvements by performing the discussed modification. By introducing more than three groups, it possibly could predict more accurately where the post typically appears in the radicalization process. RQ 4 is concerned with posts, but to answer the second goal, it should be expanded to include the classification of users, not just posts. An alternative to implementing more classes is to change the problem from a classification problem to a regression problem.

7.3 Further Discussion

This section discusses the relevant factors not discussed when answering the two thesis goals and the research questions. First, it is presented how the findings in the experiments could be used to find the point where someone leaves the radicalization process and becomes an extremist. After that, the police's definition of extremists and online surveillance are discussed, followed by the right to freedom of speech. Since the police's permissions are limited, there is a need for preventative measures, and a few are discussed in this section. Before the approaches are discussed, the psychological effects of working in this domain are presented.

7.3.1 Activation Point

The research of the master's has not defined the point where a user can be considered an extremist. If wanting to investigate the users, the results using TD-IDF and the pie diagrams can be used to define the conditions for reaching the activation point. TF-IDF scores of posts could be used to monitor users and see if the degree of the posts increased over time. Then a value could be chosen as the activation point and decided that if a user posted more than five (randomly chosen number) posts, they should be classified as extremists. Pie diagrams could be applied by investigating the distribution of a user's post over time. If the distribution changed rapidly, it could signify the user entering the radicalization process. Alternatively, it could be defined a degree of radical and/or extreme posts a user needs to have to be classified as an extremist. The changes in the use of mentions, hashtags, and URLs could help define the activation point and the changes in lengths of posts.

7.3.2 Limitations for the Police

The police's definition of extremists is restricted to someone that most likely will perform, or try to perform, an act of violence motivated by their extreme beliefs. Knowing that polarization of the society can pose a severe danger to the society and democracy, it is frightening that the police can not prevent it by using surveillance. Jilani and Smith [2019] discusses the cost of a polarized America and states that the American population increasingly segregate themselves and that political campaigns use more negativity trying to tear the opponent down instead of focusing on the party's politics. Further, they express that a polarized society increases the likelihood of violent incidents. That suggests that a single person acquiring extreme beliefs does not necessarily pose a violent threat. However, when the sum of such cases reaches a number, the danger of violence increases significantly. Therefore it could be argued that the police should be allowed to act since the long term results of the polarization indirectly do so the person in the radicalization process meets the police's criteria to be defined as an extremist.

Whom to be defined as extremists by the police differs from the definition in this thesis, and if following the criteria of the police, few, or none, of the defined extremists in the extreme dataset used in the experiments in chapter 5 would be caught or prevented from future development as extremists. The same would be the case for the radical dataset since few, or none, of them pose a violent threat. If implementing surveillance measures on the population using AI, it is likely that more potential dangerous people would be recognized, both non-violent and violent extremists, than today. Hence, with the current definition of extremists by the police, it is likely that finding extremists and preventing them from performing violent acts would be easier. Knowing that AI could help prevent dangerous situations, is it right that surveillance of the general population is not allowed? It is an enormous problem to discuss and is discussed worldwide. Article 12 in the *Universal Declaration of Human Rights* claims that every human being has a right to privacy [UN General Assembly, 1948]. Could it be argued that when the content and information are publicly available, it can not be considered private information? It is a vulnerable balance between preventing radicalization and invading someone's privacy rights and right to freedom of speech.

7.3.3 Freedom of Speech

As pointed out by the police, they are afraid of pre-censorship since the Norwegian democracy is built on the right of freedom of speech regulated by article number 19 in *Universal Declaration of Human Rights* as introduced in subsection 2.3.4. Should it be allowed to express every opinion a person may have even though it can be classified as racism or discrimination which hurts someone else? If making restrictions on the right to freedom of speech, it has to be regulated by laws according to UN General Assembly [1948]. Who should decide where the limit for illegal and legal expressions goes? If decided that the criteria for illegal are that the expression hurts someone, the allowed beliefs to express would be very limited and possible pose a more significant threat to the democracy than the current situation. If people feel that the government does not hear them due to restrictions on expression permissions, it would probably lead to mistrust in the government, which is a known indicator of people vulnerable to radicalization.

7.3.4 Preventative Measures

Preventative measures implemented in society to ensure that people do not enter the radicalization process are the best way the government can prevent a polarized society. Such measures could be to teach children to be source-critical and encourage them to know that there usually exists more than one side to a story. A person who uses time to get familiar with others' views on a case and is open to acquiring new information and possibly changing their minds would likely be more resilient to entering a radicalization process. The more information known about indicators of people vulnerable to extremism and recruitment, the easier it would be to recognize these people early and help them. A suggestion is to educate teachers to recognize these indicators since a teacher, especially in primary school, is an important person in a child's life and spends much time with them, making it easier to recognize vulnerable children. Winter et al. [2021]found that the classroom seemed to be an appropriate area for recognizing suspicious beliefs.

The average age for people being radicalized according to Klausen [2016] is 22 years. The research applies to American Al Qaida-inspired terrorists, meaning that the age of right-wing extremists can differ. However, it is reasonable to expect the age to be similar. Therefore it is unlikely that children are radicalized when attending primary school. However, they can start to express different beliefs and behave differently at an early age, which may lead to entering a radicalization process later. To notice these children and help them at an early age would be beneficial.

Educating people to be source-critical can be done at all ages, not just in primary school, but if the average radicalization age is 22 years old, it should at least be tough earlier. The education could consist of information about recognizing if friends or family are in danger of being radicalized. Informing the public about the indicators and how to report a message of concern to the authorities or other instances that can help could be useful measures. It could be to educate the worried person to be able to privately implement measures to reduce the risk of the person being radicalized or that the public takes responsibility. Kripos mentioned that they work to develop a portal for people to report concerns, so a possibility would be to make sure that the Norwegian population know about it and how to use it.

7.3.5 Psychological Effects

It has been a mental strain to be exposed to hateful expressions and extreme radical beliefs. At a point, it was decided to make a rule for how often to allow me to read extreme, hateful content. In the beginning, it was not reflected upon how it could affect the well being. However, when reading Kennedy et al. [2018] and experiencing how much capacity the read content took of the spare time, it became clear that some measures needed to be implemented. Therefore, collecting and analyzing the data took a longer time than expected. However, it was a necessary measure to prevent being too affected by the impressions.

7.4 Evaluation of Approach

This section evaluates the applied NLP and ML methods used for the experiment and suggests improvements to the approaches. Firstly, the section evaluates the choice of language features to analyze in the experiments in section 5.3. The approach of collecting data is not evaluated in this section since it was discussed when answering subsection 7.1.1. Thereafter, the NLP and ML approaches are discussed.

7.4.1 Language Features

In the experiments, it was chosen to investigate the lengths, frequent words, mentions, hashtags and URLs. Length was chosen due to previous research that disproved the hypothesis that extreme users write shorter posts than others. Therefore it was desired to check if that was only the case for Lara-Cabrera et al. [2019] and Ahmad et al. [2019] or if it also applied in the collected extreme dataset in this project. The results showed the same tendency, but it should be considered whether the results are invalid since the character limits for gabs and tweets were significantly different. Considering the frequency of words in a dataset seemed to be a helpful approach when planning the experiment. The results only confirmed the importance of the frequently used words for analysis of people in the radicalization process when comparing the frequently used words for the tree dataset in Table 6.2 (p.62).

Mentions were chosen since they can be considered a textual feature representing interaction with other users. Therefore, it was considered a valuable feature to investigate if there was a pattern of whom radical and extreme users "contacted". Hashtags were considered a way to spread information to others since it makes the post appear in specific channels, groups or search. When using hashtags, it is clear that the user wants to spread the content. Agarwal and Sureka [2015] used hashtags to create a dictionary instead of the most frequent words. That is an approach that could deliver good results in this case since the extreme users used hashtags more frequent than others. After the analysis is performed, these two features are still considered good choices for language features. Those should be further investigated and not just analyzed but used as features in the ML models.

The analysis of URLs was minimal since only the frequency of appearance in the posts was calculated and compared. Nevertheless, it is known that URLs hold much information and can contribute to conveying a message better than only text. Hence it is assumed that the frequency of URLs does not deliver a typical pattern of users in the radicalization, but investigating the content of the URLs could have a significant effect on the accuracy.

7.4.2 NLP Techniques

BoW, TF-IDF and word2vec were all used in the experiments. TF-IDF is used for calculating the degree of radical and did not deliver accurate results. The approach used in the experiment could be further developed for that purpose, but it should be investigated how to improve the accuracy, possibly combined with other methods. Word2vec includes the context in its model but was not used as a text representation model for the training of ANNs. BoW was used to transform the posts in the datasets into representation understandable for the ML approach. As known, it does not include any context information. That could be a reason why Obama, who is considered a neutral user, is classified as more radical and extreme than neutral. It seems like the ML models consider political content to be radical or extreme, but not if it is in a positive, negative or neutral context.

If the models could analyze the context of the used words, such as if it uses any of the symbols in Table 2.1 (p.13), the mentions in **??** (p.**??**) or the hashtags in Table 6.3 (p.63), the models could hopefully manage to differentiate between extreme beliefs of politics and politic content considered to be factual, not fake-news, discriminate, hateful or racist. Word2Vec is a technique that considers the context, but multiple others exist. Applying one of them as the text representation model could deliver more accurate results than BoW, for example, understanding that the politics Obama expresses are not extreme.

7.4.3 ML Approach

The experiments for the ML models only included DL, specifically ANNs. It was chosen to use ANNs since DL is a promising and growing approach applied in multiple fields. The related work in chapter 4 showed that most similar research utilized supervised ML rather than DL, which was the main reason for wanting to investigate applying ANNs for the prediction problem. Nouh et al. [2019] used TF-IDF and Word2Vec in combination with different ML approaches, and RF and ANNs delivered the best results. RF also delivered promising results for Mussiraliyeva et al. [2021] and Kursuncu et al. [2019]. Therefore it could have been better to compare different methods like RF to the ANNs to see how promising the current models are compared to others. SVM was by multiple research, as Asif et al. [2020], presented in chapter 4 as an accurate and promising ML approach for the task of predicting people in the radicalization. Comparing DL and the supervised approaches RF and SVM could have been helpful. Nevertheless, the ML was only a part of the master's thesis, not the entire task. Hence the implementation and training of all three models would have been too comprehensive.

Chapter 8

Conclusion and Future Work

The eighth chapter delivers the conclusion of the master's thesis. It presents the approach of the work and the findings. Furthermore, this chapter suggests future work based on the findings in the experiments and relevant research and analysis.

8.1 Conclusion

The research during the specialization project revealed a lack of research on recognizing right-wing radicalization on social media. This thesis analyzed differences in language use on Twitter and Gab for neutral, radical and extreme users. Extreme users were defined as users banned from Twitter. In contrast, radical users commented on a post by a known right-wing extremist but were not banned from Twitter, and the neutral users were presented in an open-source dataset. The analysis found that extreme users tend to use hashtags more frequently than neutral users and that the frequently used hashtags and words by the extreme and radical users were more political-related than for neutral users. The frequently used words in the radical dataset contained fewer political words than in the extreme dataset and more than in the neutral dataset. Additionally, the post length and URL use were analyzed. Language use for extremists, radicals and neutral users is expected to be a promising feature for recognizing vulnerable social media users.

A variant of TF-IDF valuing frequently used words calculated the degree of how extreme a post is. The results were not satisfactory since the post length mattered more than desired. Word2vec found similarities between words and should be tried as a text representation model for ANNs to capture the context. BoW, which does not consider the context, was used for text representation for the ANNs. Neutral, radical and extreme users were visualized in pie diagrams by classifying their posts. The results suggested that political content is an indicator of extreme content. Including context in training is expected to improve the model's ability to distinguish between objective and radical use of political content in a post. Minimizing the number of neutral posts classified as radical or extreme was the most critical evaluation for the experiment. The best model given based on that condition gave a recall of 86.6% for the neutral class and total accuracy of 50.3% when predicting labels of posts in the prediction dataset. The training and validation accuracies were between 81% and 84.5%.

It is crucial to define extremism since multiple definitions exist, and the treatment of people meeting different criteria should be different. Therefore it is important to have an appropriate dataset for the chosen definition, which suggests that the data in the experiments should have been validated to improve the accuracy of the results.

8.2 Future Work

This thesis has delivered valuable results and discussions to the field of right-wing extremism. Two novel datasets and a dictionary of right-wing words were delivered, in addition to a modified TF-IDF model and ANN models for multi-label classification. The mentioned contributions were helpful but can be modified and improved. In this section, it is suggested future work that could improve the results. Possible tasks that are motivated by the experiments are presented at the end.

8.2.1 Datasets

Throughout the thesis, it has been expressed that it would be desired to use data from one platform. Therefore, the first task should be collecting and annotating data from a single platform, preferably Gab, representing extreme, radical and neutral data. Alternatively, just extreme and neutral data if using a regression approach like suggested in subsection 8.2.4. Further, the collected data should be annotated thoroughly to minimize the number of wrongly classified posts since extreme users probably post neutral posts, and neutral users may occasionally post extreme content. It is expected that it would significantly increase the performance of all the other experiments without changing the approaches.

8.2.2 Analysis

It is possible to investigate other language features like the use of symbols, verbs, or swearwords. Nevertheless, the text analysis performed delivered enough results to suggest that for further analysis, the language features in this task would be a good starting point when using this thesis as the basis. The ML approaches should then be tested with these features, and then the results should be compared to the current results. Could hashtags and mentions more accurate predict vulnerable users than using BoW for the whole corpus?

8.2.3 A Post's Degree of Extreme

The radicalization process happens over time, not on a single occasion, making it likely that the expression behavior of users in the process is continuous. That suggests that calculating the degree of how extreme a post is could make it easier to see where in the process a user is than if only classifying into three groups. As mentioned in earlier chapters, it should be tried to improve the TF-IDF method for calculating how extreme posts are. Other methods should be tried and compared to TF-IDF.

8.2.4 Classification and Regression

This thesis wanted to classify users into one of three categories, which may be too few when wanting to recognize people early in the radicalization process. The current solution may result in recognizing these people later than desired. One solution is to have more than three categories. Alternatively, the problem could be transformed into a regression problem. Then it could be predicted how extreme a post is using ML, not only NLP. Further, the classification or regression should be expanded to perform the same process on users, not just posts.

8.2.5 Activation Point

When having methods to figure out where in the radicalization process a user is, it should be investigated how to decide where the user becomes an extremist in the process. If the point can be defined, it could be used to find the critical period where preventative measures must be implemented before it is too late.

8.2.6 Different Language

The experiments were solely based on English posts. The extreme content varies between countries, and the content is written in different languages. Therefore it could be helpful to adjust the models to other languages like Norwegian. The

Norwegian police want similar research on Norwegian social media content. It would be not only an exciting task but also a meaningful and needed task.

Bibliography

ADL (2021). Stormfront. https://www.adl.org/education/references/hate-symbols/stormfront (Accessed: 08.12.21).

Agarwal, S. and Sureka, A. (2015). *Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter.*

Ahmad, S., Asghar, D. M., Alotaibi, F., and Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences*, 9:24.

Åkerlund, M. (2021). Dog whistling far-right code words: the case of âculture enricher' on the swedish web. *Information, Communication & Society*, 0(0):1–18.

Al-Saggaf, Y. (2018). Online radicalisation along a continuum: From when individuals express grievances to when they transition into extremism. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, volume 255, pages 429–440.

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., and Moustafa, A. A. (2019). The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Frontiers in Computational Neuroscience*, 13:31.

Aleroud, A., Abu-Alsheeh, N., and Al-Shawakfa, E. (2020). A graph proximity feature augmentation approach for identifying accounts of terrorists on Twitter. *Computers and Security*, 99.

Amnesty International Norway (2019). 21 politiske tiltak mot netthets. https://amnesty.no/sites/default/files/vedlegg/21-tiltak-amnesty-ferdigstilt.pdf.

Amnesty International UK (2020). What is freedom of speech? https://www.amnesty.org.uk/free-speech-freedom-expression-human-right (Accessed: 12.05.22).

Angus, C. (2016). Radicalisation and violent extremism: causes and responses. https://www.parliament.nsw.gov.au/researchpapers/Documents/radicalisation-and-violent-extremism-causes-and-/Radicalisation%20eBrief.pdf.

Arya, D., Rudinac, S., and Worring, M. (2019). Predicting Behavioural Patterns in Discussion Forums using Deep Learning on Hypergraphs. In *Proceedings - International Workshop on Content-Based Multimedia Indexing*, volume 2019-September.

Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., and Shah, J. (2020). Sentiment analysis of extremism in social media from textual information. *Telematics and Informatics*, 48.

Barhamgi, M., Masmoudi, A., Lara-Cabrera, R., and Camacho, D. (2018). Social networks data analysis with semantics: application to the radicalization problem. *Journal of Ambient Intelligence and Humanized Computing*.

Beheshti, A., Moraveji-Hashemi, V., Yakhchi, S., Motahari-Nezhad, H. R., Ghafari, S. M., and Yang, J. (2020). Personality2Vec: Enabling the analysis of behavioral disorders in social networks. In *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 825–828.

Behr, I., Reding, A., Edwards, C., and Gibbion, L. (2013). Radicalisation in the Digital Era. Report, RAND Europe.

Bengfort, B., Bilbro, R., and Ojeda, T. (2018). *Applied Text Analysis with Python*. O'Reilly Media, Inc.

Benigni, M., Joseph, K., and Carley, K. (2017). Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PLOS ONE*, 12:e0181405.

Berger, J. M. and Bill, S. (2013). Who matters online: measuring influence, evaluating content and countering violent extremism in online social networks.

Borum, R. (2003). Understanding the Terrorist Mindset. *FBI Law Enforcement Bulletin*, 72.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Britannica Dictionary (2022). Britannica Dictionary definition of NEOâNAZI. https://www.britannica.com/dictionary/neo-nazi (Accessed: 29.04.22).

Brownlee, J. (2016). Naive Bayes for Machine Learning. https://machinelearningmastery.com/naive-bayes-for-machine-learning/ (Accessed: 10.12.21). Machine Learning Mastery.

Brownlee, J. (2019). A Gentle Introduction to the Rectified Linear Unit (ReLU). https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ (Accessed: 02.06.22).

BuiltIn (2021). Introduction to AI. https://builtin.com/artificial-intelligence (Accessed: 08.11.21).

Bundesamtür Verfassungsschutz (2018). Right-wing extremism: Signs, symbols and banned organisations. Available at: https://www.verfassungsschutz.de/SharedDocs/publikationen/EN/right-wing-extremism/2018-10-right-wing-extremism-symbols-and-organisations.pdf?__blob=publicationFile&v=10).

Cardenas, P., Obara, B., Theodoropoulos, G., and Kureshi, I. (2018). Defining an Alert Mechanism for Detecting likely threats to National Security. In *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pages 1575–1580.

Cutler, A., Cutler, D., and Stevens, J. (2011). *Random Forests*, volume 45, pages 157–176.

Daniels, J. (2018). The algorithmic rise of the âalt-rightâ. *Contexts*, 17(1):60–65.

Dollarhide, M. (2021). Social Media. https://www.investopedia.com/terms/s/social-media.asp (Accessed: 15.11.21). Investopedia.

Facebook, I. (2021). Facebook Reports Third Quarter 2021 Results. https://s21.q4cdn.com/399680738/files/doc_news/Facebook-Reports-Third-Quarter-2021-Results-2021.pdf.

Fernandez, M., Asif, M., and Alani, H. (2018). Understanding the roots of radicalisation on Twitter. In *WebSci 2018 - Proceedings of the 10th ACM Conference on Web Science*, pages 1–10.

Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A., and Galstyan, A. (2016). *Predicting Online Extremism, Content Adopters, and Interaction Reciprocity.*

Fraštíková, S. and Demčišák, J. (2019). The Language of the Right-wing Populism: A Lexical Analysis of the Texts by the Freedom Party of Austria. *ACCS2019 Conference Proceedings.*

Gaikwad, M., Ahirrao, S., Phansalkar, S., and Kotecha, K. (2021). Online Extremism Detection: A Systematic Literature Review with Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools. *IEEE Access*, 9:48364–48404.

Garvik, O. and Stenersen, A. (2018). Profetens Ummah. `https://snl.no/Profetens_Ummah` (Accessed: 06.06.22). Store norske leksikon (SNL).

Giordano, D. (2020). 7 tips to choose the best optimizer. `https://towardsdatascience.com/7-tips-to-choose-the-best-optimizer-47bb9c1219e` (Accessed: 02.06.22). Towards Data Science.

Haney-López, I. (2014). *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.

Hervik, P. (2019). Ritualized Opposition in Danish Online Practices of Extremist Language and Thought. *International Journal of Communication (19328036)*, 13.

HM Government (2015). Counter-extremism strategy. Command paper, HM Government. Presented to Parliament by the Secretary of State for the Home Department by Command of Her Majesty.

IACP, C. o. T. (2012). A common lexicon. Countering Violent Extremism (CVE) Working Group.

IBM (2020). Natural Language Processing (NLP). `https://www.ibm.com/cloud/learn/natural-language-processing` (Accessed: 05.11.21). IMB Cloud Education.

IBM Cloud Education (2020a). Deep Learning. `https://www.ibm.com/cloud/learn/deep-learning` (Accessed: 10.12.21).

IBM Cloud Education (2020b). Neural Networks. `https://www.ibm.com/cloud/learn/neural-networks` (Accessed: 10.12.21).

International Association of Chiefs of Police (2014). Online Radicalization to Violent Extremism. Awareness brief, Washington, DC: Office of Community Oriented Policing Services.

Jilani, Z. and Smith, J. A. (2019). What Is the True Cost of Polarization in America? `https://greatergood.berkeley.edu/article/item/what_is_the_true_cost_of_polarization_in_america` (Accessed: 21.05.22).

Jones, S. G., Doxsee, C., and Harrington, N. (2020). The Escalating Terrorism Problem in the United States. Csis briefs, Center for Strategic and International Studies.

Justis- og beredskapsdepartementet (2020). Handlingsplan mot radikalisering og voldelig ekstremisme - revisjon 2020. Originally published by the Ministry of Culture.

Kemp, S. (2020). Digital 2021 October Global Statshot Report. Technical report, DataReportal. `https://datareportal.com/reports/digital-2021-october-global-statshot` (Accessed: 02.11.21).

Kemp, S. (2022). Digital 2022 April Global Statshot Report. Technical report, DataReportal. `https://datareportal.com/reports/digital-2022-april-global-statshot` (Accessed: 14.05.22).

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., and et al. (2018). Introducing the gab hate corpus: Defining and applying hate-based rhetoric to social media posts at scale.

Keras (2022). Losses. `https://keras.io/api/losses/` (Accessed: 30.04.22).

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*.

Klausen, J. (2016). A behavioral study of the radicalization trajectories of american 'homegrown' al qaeda-inspired terrorist offenders. Funded by the U.S. Department of Justice.

Knudsen, R. A. (2020). Measuring radicalisation: risk assessment conceptualisations and practice in England and Wales. *Behavioral Sciences of Terrorism and Political Aggression*, 12(1):37–54.

Koech, K. E. (2020). Softmax Activation Function â How It Actually Works. `https://towardsdatascience.com/softmax-activation-function-how-it-actually-works-d292d335bd78` (Accessed: 02.06.22). Towards Data Science.

Kofod-Petersen, A. (2018). How to do a Structured Literature Review in computer science. Technical Report version 2, Department of Computer Science, Norwegian University of Science and Technology Science.

Kursuncu, U., Gaur, M., Castillo, C., Alambo, A., Thirunarayan, K., Shalin, V., Achilov, D., Arpinar, I. B., and Sheth, A. (2019). Modeling Islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Lara-Cabrera, R., Gonzalez-Pardo, A., and Camacho, D. (2019). Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter. *Future Generation Computer Systems*, 93:971–978.

Legislation.gov.uk (2021). Terrorism Act 2000. `https://www.legislation.gov.uk/ukpga/2000/11/section/1` (Accessed: 31.10.21).

Lexalytics (2021). Sentiment Analysis Explained. `https://www.lexalytics.com/technology/sentiment-analysis` (Accessed: 10.12.21).

López-Sáncez, D., Revuelta, J., de la Prieta, F., and Corchado, J. M. (2018). Towards the automatic identification and monitoring of radicalization activities in Twitter. In *Communications in Computer and Information Science*, volume 877, pages 589–599.

Marwick, A. E. and Lewis, R. (2017). Media manipulation and disinformation online.

Masood, M. A. and Abbasi, R. A. (2021). Using graph embedding and machine learning to identify rebels on Twitter. *Journal of Informetrics*, 15(1).

Mbaabu, O. (2020). Introduction to Random Forest in Machine Learning. `https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/` (Accessed: 10.12.21). Section.

Meta Platforms, Inc (2022). Meta Reports First Quarter 2022 Results. `https://s21.q4cdn.com/399680738/files/doc_financials/2022/q1/Meta-03.31.2022-Exhibit-99.1_Final.pdf`.

Milaĉić, F. (2021). The Negative Impact of Polarization on Democracy. Available at: `http://library.fes.de/pdf-files/bueros/wien/18175.pdf`).

Moghaddam, F. (2005). The staircase to terrorism a psychological exploration. *American Psychologist*, 60(2):161–169.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning.* The MIT Press.

Munk, T. B. (2017). 100,000 false positives for every real terrorist: Why anti-terror algorithms don't work. *First Monday*, 22(9).

Mussiraliyeva, S., Bolatbek, M., Omarov, B., Medetbek, Z., Baispay, G., and Ospanov, R. (2020). On detecting online radicalization and extremism using natural language processing. In *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*.

Mussiraliyeva, S., Omarov, B., Bolatbek, M., Ospanov, R., Baispay, G., Medetbek, Z., and Yeltay, Z. (2021). Applying Deep Learning for Extremism Detection. In *Communications in Computer and Information Science*, volume 1393, pages 597–605.

Nouh, M., Jason Nurse, R. C., and Goldsmith, M. (2019). Understanding the radical mind: Identifying signals to detect extremist content on Twitter. In *2019 IEEE International Conference on Intelligence and Security Informatics, ISI 2019*, pages 98–103.

Oslo politidistrikt (2019). HATKRIMINALITET Anmeldt hatkriminalitet 2019. `https://www.politiet.no/globalassets/dokumenter/oslo/rapporter/anmeldt-hatkriminalitet-oslo/anmeldt-hatkriminalitet-i-oslo-politidistrikt-2021.pdf`.

Oslo politidistrikt (2022). Anmeldt hatkriminalitet i Oslo politidistrikt 2021. `https://www.politiet.no/globalassets/dokumenter/oslo/rapporter/anmeldt-hatkriminalitet-oslo/anmeldt-hatkriminalitet-i-oslo-politidistrikt-2021.pdf`.

Oxford Learner's Dictionary (2021). Definition of extremism noun from the Oxford Advanced Learner's Dictionary. `https://www.oxfordlearnersdictionaries.com/definition/english/extremism` (Accessed: 28.10.21).

Pew Research Center (2014). Political Polarization in the American Public.

Politiet (2021). Hatkriminalitet - Anmeldt hatkriminalitet i 2020. `https://www.politiet.no/globalassets/dokumenter/oslo/rapporter/anmeldt-hatkriminalitet-oslo/anmeldt-hatkriminalitet-i-oslo-2020.pdf`.

Politiet (2022). Politiets trusselvurdering 2022. `https://www.politiet.no/globalassets/04-aktuelt-tall-og-fakta/politiets-trusselvurdering-ptv/politiets-trusselvurdering-2022.pdf`.

Politiet Kripos (2022). Kripos' strategi 2022-2025. `https://www.politiet.no/globalassets/dokumenter/kripos/kripos-strategi-2022-2025.pdf`.

PST (2019). Temarapport: Hvilken bakgrunn har personer i høyreekstreme miljøer i norge? *Politiets sikkerhetstjeneste.* https://www.pst.no/globalassets/artikler/utgivelser/temarapport_pst_-hvilken-bakgrunn-har-personer-i-hoyreekstreme-miljoer-i-norge.pdf (Accessed: 24.03.22).

PST (2022). National Threat Assessment for 2022. Available at: https://www.pst.no/alle-artikler/trusselvurderinger/ntv-2022/.

Ravndal, J. A., Lygren, S., Jupskås, A. R., and Bjørgo, T. (2020). RTV Trend Report 2020 - Right-Wing Terrorism and Violence in Western Europe, 1990 - 2019. Available at: https://www.sv.uio.no/c-rex/english/publications/c-rex-reports/2020/rtv-trend-report/c-rex-rtv-trend-report-2020.pdf).

Regjeringen.no (2019). Hva er radikalisering og voldelig ekstremisme? https://www.regjeringen.no/no/sub/radikalisering/om-forebyggende-arbeid/hva-er-radikalisering-og-voldelig-ekstre/id663761/ (Accessed: 31.10.21).

Rehman, U., Abbas, S., Khan, M. A., Mustafa, G., Fayyaz, H., Hanif, M., and Saeed, M. A. (2020). Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning. *Computers, Materials and Continua*, 66(2):1075–1090.

Rowe, M. and Saif, H. (2016). Mining Pro-ISIS Radicalisation Signals from Social Media Users. *ICWSM-16: 10th International AAAI Conference on Web and Social Media.*

Russel, S. and Norvig, P. (2016). *Artificial Intelligence: A Modern Approach.* Pearson. Third Edition.

Santos, M. (2020). Precision or Recall: Which Should You Use? *Towards Data Science.*

Schmid, D. A. P. (2013). Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review. *ICCT Research Paper.*

Secretary of State for the Home Department (2018). CONTEST - The United Kingdom's Strategy for Countering Terrorism. Presented to Parliament by the Secretary of State for the Home Department by Command of Her Majesty.

Singh, P. (2019). Fundamentals of Bag of Words and TF-IDF. https://medium.com/analytics-vidhya/fundamentals-of-bag-of-words-and-tf-idf-9846d301ff22 (Accessed: 20.04.22).

Skleparis, D. and Knudsen, R. A. (2020). Localising 'radicalisation': Risk assessment practices in Greece and the United Kingdom. *British Journal of Politics and International Relations*, 22(2):309–327.

Tan, S., Guan, Z., Cai, D., Qin, X., Bu, J., and Chen, C. (2014). Mapping users across networks by manifold alignment on hypergraph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

TensorFlow (2022). tf.keras.losses.sparsecategoricalcrossentropy. `https://www.tensorflow.org/api_docs/python/tf/keras/losses/SparseCategoricalCrossentropy` (Accessed: 30.04.22).

The Norwegian Government (2020). International efforts to promote freedom of expression and independent media. `https://www.regjeringen.no/en/topics/foreign-affairs/human-rights/ny-struktur/promote_freedom/id2358336/` (Accessed: 12.05.22).

Tundis, A., Bhatia, G., Jain, A., and Muhlhauser, M. (2018). Supporting the identification and the assessment of suspicious users on Twitter social media. In *NCA 2018 - 2018 IEEE 17th International Symposium on Network Computing and Applications*.

Twitter (2021). Counting characters. `https://developer.twitter.com/en/docs/counting-characters` (Accessed: 03.11.21). Developer Platform.

Twitter, Inc (2022). Twitter Announces First Quarter 2022 Results. `https://s22.q4cdn.com/826641620/files/doc_financials/2022/q1/Final-Q1'22-earnings-release.pdf`.

@TwitterIR (2021). Q3 2021 Letter to Shareholders. `https://s22.q4cdn.com/826641620/files/doc_financials/2021/q3/Final-Q3'21-Shareholder-letter.pdf`.

Uc-Cetina, V., Navarro-Guerrero, N., Martin-Gonzalez, A., Weber, C., and Wermter, S. (2021). Survey on reinforcement learning for language processing.

Udanor, C. and Anyanwu, C. C. (2019). Combating the challenges of social media hate speech in a polarized society: A Twitter ego lexalytics approach. *Data Technologies and Applications*, 53(4):501–527.

UN General Assembly (1948). Universal Declaration of Human Rights. Available at: `https://www.refworld.org/docid/3ae6b3712c.html` (Accessed: 12.05.22).

Winter, C., Heath-Kelly, C., Kaleem, A., and Mills, C. (2021). A moral education? British Values, colour-blindness, and preventing terrorism. *Critical Social Policy.*

Wolfowicz, M., Perry, S., Hasisi, B., and Weisburd, D. (2021). Faces of radicalism: Differentiating between violent and non-violent radicals by their social media profiles. *Computers in Human Behavior*, 116.

Xu, F., Sun, D., Li, Z., and Li, B. (2017). Research on online supporting community of extreme organization by AI-SNA based method. In *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, volume 2017-November, pages 546–551.

Yadav, P. (2018). Decision Tree in Machine Learning. https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96 (Accessed: 10.12.21). Towards data science.

Zote, J. (2021). How long should social posts be? Try this social media character counter. https://www.investopedia.com/terms/s/social-media.asp (Accessed: 03.11.21).

Appendix

A The Penal Code

A.1 §77: Aggravating circumstances

In connection with sentencing, aggravating factors to be given particular consideration are that the offence:

a) *was committed by means or methods which are particularly dangerous or carry a considerable potential for harm,*

b) *placed human life or health at risk or caused loss of welfare,*

c) *was intended to have a substantially more serious outcome or this could easily have been the consequence,*

d) *was committed in a particularly reckless manner,*

e) *formed part of a planned or organised enterprise,*

f) *was committed by multiple persons acting together,*

g) *was perpetrated by the offender exploiting or misguiding young persons, persons in a very difficult life situation, who are mentally disabled or in a dependent relationship with the offender,*

h) *affected persons who are defenceless or particularly vulnerable to criminal offences,*

i) *was motivated by a person's religion or life stance, skin colour, national or ethnic origin, homosexual orientation, disability or other circumstances relating to groups with a particular need for protection,*

j) *was committed in the course of public service or was perpetrated by violating a special trust,*

k) *was committed by a person who has previously been the subject of a criminal sanction for similar acts or other acts of relevance to the case,* *l) was committed in the presence of a child under 15 years of age.*

The Norwegian version can be found here: https://lovdata.no/lov/2005-05-20-28/§77 ad the English version can be found here: https://lovdata.no/NLE/lov/2005-05-20-28/§77.

A.2 §185: Hate speech

A penalty of a fine or imprisonment for a term not exceeding three years shall be applied to any person who with intent or gross negligence publicly makes a discriminatory or hateful statement. Â«StatementÂ» includes the use of symbols. Any person who in the presence of others, with intent or gross negligence, makes such a statement to a person affected by it, see the second paragraph, is liable to a penalty of a fine or imprisonment for a term not exceeding one year. Â«Discriminatory or hateful statementÂ» means threatening or insulting a person or promoting hate of, persecution of or contempt for another person based on his or her

a) skin colour or national or ethnic origin,

b) religion or life stance,

c) homosexual orientation, or

d) reduced functional capacity.

The Norwegian version can be found here: https://lovdata.no/lov/2005-05-20-28/§185 and the English version can be found here: https://lovdata.no/NLE/lov/2005-05-20-28/§185.

A.3 §18: Self-defence

An act which would otherwise be punishable, is lawful when it

a) is committed to avert an unlawful attack,

b) does not exceed what is necessary, and

c) does not clearly go beyond what is justifiable, taking into account the dangerousness of the attack, the type of interest the attack violates, and the culpability of the assailant.

The rule in the first paragraph applies correspondingly to any person who effects a lawful arrest or attempts to prevent a person from evading being remanded in custody or serving a custodial sentence.

The exercise of public authority may only be met with an act of self-defence if the exercise of authority is unlawful and the person who exercises it acts with intent or gross negligence.

The Norwegian version can be found here: https://lovdata.no/lov/2005-05-20-28/§18 ad the English version can be found here: https://lovdata.no/NLE/lov/2005-05-20-28/§18.

B A Retrieved JSON object from Gab

```
"1": {
    "id": "1",
    "created_at": "2022-05-02T15:30:29.386Z",
    "revised_at": null,
    "in_reply_to_id": null,
    "in_reply_to_account_id": null,
    "sensitive": false,
    "spoiler_text": "",
    "visibility": "public",
    "language": "en",
    "uri": "/example_user//1",
    "url": "https://gab.com/example_user/1",
    "direct_replies_count": 0,
    "replies_count": 3,
    "reblogs_count": 2,
    "pinnable": false,
    "pinnable_by_group": false,
    "favourites_count": 10,
    "quote_of_id": null,
    "expires_at": null,
    "has_quote": false,
    "bookmark_collection_id": null,
    "quotes_count": 2,
    "favourited": false,
    "reblogged": false,
    "muted": false,
    "content": "This is an example post",
    "rich_content": "",
    "plain_markdown": null,
    "reblog": null,
    "quote": null,
    "account": {
      "id": "1",
      "username": "example_user",
      "acct": "example_user",
      "display_name": "Example User",
      "locked": false,
      "bot": false,
      "created_at": "2022-01-01T18:05:14.968Z",
      "note": "<p>I am an example of a Gab user</p>",
      "url": "https://gab.com/example_user",
      "avatar": "profile_picture_link",
      "avatar_static": "profile_picture_link",
      "avatar_small": "profile_picture_link",
      "avatar_static_small": "profile_picture_link",
      "header": "null,
      "header_static": "null",
      "is_spam": false,
      "followers_count": 193,
      "following_count": 81,
      "statuses_count": 3,
      "is_pro": false,
      "is_verified": false,
      "is_donor": false,
      "is_investor": false,
      "show_pro_life": false,
      "emojis": [],
      "fields": []
    },
    "group": null,
    "media_attachments": [],
    "mentions": [@example2],
    "tags": [],
    "emojis": [],
    "card": null,
    "poll": null,
    "body": "This is an example post"
  }
}
```

Listing 1: An example of a post in JSON file.

C Literature Review

In this appendix section, the approach used to retrieve the literature from the previous semester is presented. This section is copied from the specialization project with only minor changes.

C.1 Planning

The planning phase consists of five steps. Steps 1-3 are performed in previous chapters. Hence, this section will focus on 4 and 5.

Five steps of planning:

1. Identification of the need for a review
2. Commissioning a review
3. Specifying the research question(s)
4. Developing a review protocol
5. Evaluating the review protocol

It is essential to conduct a review protocol to ensure that each step is done correctly and can be reproduced. The review protocol is continuously updated in order to optimize it. Then both steps 4 and 5 are completed. This section explains how the review is performed.

C.2 Conducting

The next phase must define how to retrieve relevant literature for the research questions and extract the relevant information from that literature. The five steps of the phase are shown below. At first, key terms need to be defined to conduct a query for searching through the chosen source. Thereafter, the retrieved literature has to be reduced by introducing some criteria. Finally, data has to be extracted and analyzed from the remaining literature.

1. Identification of research
2. Selection of primary studies
3. Study quality assessment
4. Data extraction and monitoring
5. Data synthesis

	Group 1	Group 2	Group 3	Group 4
Term 1	*Extremism*	*Social Media*	*Prevention*	*Artificial Intelligence*
Term 2	*Radicalism*	*Twitter*	*Identifying*	*Machine Learning*
Term 3	*Terrorism*	*Facebook*	*Detection*	*Prediction*
Term 4	*Right-Wing*	*Gab*		*Natural Language Processing*

Table 1: Search terms.

Step 1: Identification of research

The first step aims to retrieve relevant literature based on key terms, also referred to as search terms, used to search for in the chosen source.

Source
Scopus[1] was used as search database. It was chosen since it contains peer-reviewed literature. The website also allows for multiple filters, making finding relevant literature easy.

Search terms
The search terms are presented in Table 1. Group 1 is included to make sure the retrieved literature is related to extremism regardless of which variant of the word is used. The second group ensures that they discuss extremism on social media. Further, group 3 is included because this paper aims to identify users before they are radicalized. Therefore it is essential to review literature that has investigated how to identify users prior to the activation point of becoming extremists. The last group retrieves papers that try to solve the problems using artificial intelligence.

Query
Based on the search terms, a query was conducted to retrieve relevant literature as displayed in Equation 1. The query was run on Scopus. After execution, this equation returned results relevant for all four groups.

$$\begin{gathered}(Extremism \lor Radicalism \lor Terrorism \lor Right - Wing) \land \\ (Social\ Media \lor Twitter \lor Facebook \lor Gab) \land \\ (Prevention \lor Identifying \lor Detection) \land \\ (Artificial\ Intelligence \lor Machine\ Learning \lor \\ Prediction \lor NaturalLanguageProcessing)\end{gathered} \tag{1}$$

[1] https://www.scopus.com/

Result
The query returned 1165 results to bring to step 2.

Step 2: Selection of primary studies

Documents from before 2018 were removed to narrow the scope. Since the field is constantly evolving, it was decided that the newest documents most likely would be the most relevant. Despite the filtration in 2018, some older articles were retrieved, probably due to the publication date in a specific paper. Since they only were a few years older, they were kept. This reduces the result to 801 documents. Next, the filter *Computer Science* was applied. This reduced the result to 383 documents. The results were sorted by relevance for the query.

Step 3: Study quality assessment

The protocol for concluding the final articles contains inclusion criteria (IC) and quality criteria (QC). IC consists of primary and secondary criteria. The literature passing step 2 was evaluated using the criteria in Table 2. Not all literature not fulfilling the criteria was removed due to the perception of relevance.

Criteria Identification	Criteria
IC1	The study's focus is extremism on social media.
IC2	The study uses AI.
IC3	The study investigates people vulnerable to extremism or the radicalisation process.
IC4	The study investigates textual features.
QC1	It is clear what the aim of the study is.
QC2	The paper puts its study into context with other similar research.

Table 2: Inclusion (IC) and quality (QC) criteria.

Primarily inclusion criteria (IC)
Each of the 383 documents was reviewed based on its title and abstract. The ones that did not meet the primary IC or were perceived as irrelevant were removed. Eighty-four documents advanced to the next stage.

Secondary inclusion criteria (IC)
Each of the 84 documents was reviewed based on the secondary criteria screening

the full text. The ones that did not meet the secondary IC or required paid access were removed.

Quality criteria (QC)
The relevant ones were reviewed by reading the whole text and evaluating them using $Q1$ and $Q2$. A total of 21 articles were chosen as the final relevant literature.

Final quality criteria
Then new quality criteria in Table 3 were created to supplement the ones in Table 2, to exclude more literature. Each article was given a score based on the QC displayed in Table 4. None of the literature was removed based on their score, but it was discussed with the perceived relevance based on its score in mind. Each criterion was given a max value of 1, except for QC8, which could be 2. QC8 was considered the most important since the discussion of the time prior to radicalization is the paper's focus.

Criteria	identification
QC1	Is there a clear presentation of the aim of the research?
QC2	Does the research include previous work done by others?
QC3	Is it used a reasonable dataset?
QC4	Is the approach chosen suitable for the problem?
QC5	Is the results evaluated using accuracy metrics?
QC6	Does discussion answer the aim of the research?
QC7	Include future work?
QC8	Does it address the users prior to the radicalization point?
QC9	Apply ML to achieve their goal?
QC10	Is the research concerned with NLP?

Table 3: The quality assessment criteria.

The scores are subjective and might be given differently by someone else.

Step 4: Data extraction and monitoring

In step 4, nine types of data were chosen to retrieve from the literature. For each of the articles, these nine features were extracted and presented in Table 5.

1. **Title**

2. **Author**

3. **Year**

4. **Psychological or social theories**

5. **NLP elements**

6. **Algorithm(s)**

7. **Dataset**

8. **Extremism**

9. **Analysis, prediction or identification**
   The category given for an article is based on the perception when reading it. It is given based on the part of the article that is considered most relevant for the task. Hence someone else might have classified it differently.

Literature	QC1	QC2	QC3	QC4	QC5	QC6	QC7	QC8	QC9	Q10	Total
1. [Asif et al., 2020]	1	1	1	1	1	1	1/4	1	1	1	9.25
2. [López-Sáncez et al., 2018]	1	1/2	1/2	1	0	3/4	1	2	0	0	6.75
3. [Mussiraliyeva et al., 2021]	1	0	1/4	1/2	1	1/2	1	0	1	1/4	5.5
4. [Mussiraliyeva et al., 2020]	1	1	1/2	1	1	1/2	1	0	1	1	8
5. [Arya et al., 2019]	1/2	1/2	1	1	3/4	1 /2	0	1	1/4	1	6.5
6. [Aleroud et al., 2020]	1	1	1	1	1	1	1	0	1	1	9
7. [Kursuncu et al., 2019]	1	1	1	1	1	1	1	1	1	1	10
8. [Masood and Abbasi, 2021]	1	1	1	1	1	1	1/4	0	1	1	8.25
9. [Fernandez et al., 2018]	1	1	1	1	1	1	1	2	1	1/2	10.5
10. [Wolfowicz et al., 2021]	3/4	1	1	1	1	1	3/4	0	1	0	7.5
11. [Al-Saggaf, 2018]	1	1	1	1	0	1	3/4	2	1/2	1/2	8.75
12. [Nouh et al., 2019]	1	1	1	1	1	1	1	0	1	1	9
13. [Barhamgi et al., 2018]	1	3/4	3/4	1	3/4	1	1	1	1/2	0	7.75
14. [Xu et al., 2017]	1	3/4	1	1	0	1	1	0	1/2	1	7.25
15. [Rowe and Saif, 2016]	1	1	1	1	1/2	1	1	2	1/4	1	9.75
16. [Tundis et al., 2018]	1	1	1/2	3/4	1	1	1	1	0	1	8.25
17. [Beheshti et al., 2020]	1	1/2	3/4	1	0	3/4	1	1	1	1	7
18. [Cardenas et al., 2018]	1	1	1	1	1/2	1	1	1/2	1	1/4	8.25
19. [Gaikwad et al., 2021]	1	1	0	1	0	1	1	0	1	1	7
20. [Udanor and Anyanwu, 2019]	1	1	1/2	1	1/2	1	1	0	1/2	3/4	7.5
21. [Rehman et al., 2020]	1	1	1	1	1	1	1	0	1	1	9
22. [Ferrara et al., 2016]	1	1	1	1	1	1	1	0	1	0	8
23. [Agarwal and Sureka, 2015]	1	1/2	3/4	1	1	1	0	0	1	1/2	6.75
24. [Benigni et al., 2017]	1	1	3/4	1	1	1	1/2	1/2	1	0	7.75
25. [Lara-Cabrera et al., 2019]	1	1	3/4	1	0	1	1	2	0	1/2	8.25

Table 4: Quality assessment score.

Step 5: Data synthesis

The important results of the data synthesis are presented in chapter 4.

D Retrieved Literature

All tables in this section are copied from the specialization project.

ID	Title	Author	Year	Psychological or social theories	NLP Elements	Algorithm(s)	DataSet	Extremism	Predict, Identify, Analysis
1	Sentiment analysis of extremism in social media form textual information	Asif et al.	2020	—	Sentimental analysis, TF-IDF	Multinomial NB, Linear Support Vector Classifier	Conducted themselves from Facebook News Pages	Radicals writing Urdu, English	I, P
2	Towards the Automatic Identification and Monitoring of Radicalization Activities in Twitter	López-Sáncez et al.	2018	—	—	Will be part of future work	Downloading data from Twitter	Right-Wing	I, P
3	Applying Deep Learning for Extremism Detection	Mussiraliyeva et al.	2021	—	Preprocessing	CNN, LSTM	Not specified	Radicas writing Kazakh	I
4	On Detecting Online Radicalization and Extremism Using Natural Language Processing	Mussiraliyeva et al.	2020	—	StringToWordVector, Word2Vec, TF-IDF	Gradient boosting, Random Forest	Gathered Data from VKontakte	Right-Wing	I
5	Predicting Behavioural Patterns in Discussion Forums using Deep Learning on Hypergraphs	Arya et al.	2019	—	Multimedia	MGCNN	Gathered from Stormfront	White nationalists, suprememacists, neo-Nazi	P
6	A graph proximity feature augmentation approach for identifying accounts of terrorists on Twitter	Aleroud et al.	2020	—	TD-IDF, TM, TC, NOV	SVM, kNN, DT, RF	Two Twitter data sets from Kaggle	ISIS	I
7	Modelling Islamist Extremist Communications on Social Media using Contextual Dimentions: Religion, Ideology, and Hate	Kursunch et al.	2019	Not theory but textual dimensions: Relogion, ideology, hate	Word2Vec, n-gram	RF, NB	Two pro-ISIS twitter datasets and one anti-ISIS	ISIS	P
8	Using graph embedding and machine learning to identify rebels on twitter	Masood and Abbasi	2021		StemWord, removeStopWordNodes, graph2vec	SVM, RF, GNB, LR	Tweets from rebels, counter rebels and normal users form five countries	Rebels	I
9	Understanding the Roots of Radicalisation on Twitter	Fernandez et al.	2018	Roots of Radicalization	n-grams	J48, NB, LR	Two Twitter datasets frim Kaggle: pro-ISIS & not pro-ISIS	ISIS	P
10	Faces of radicalism: Differentiating between violent and non-violent radicals by their social media profiles	Wolfowicz et al.	2020	Social learning theory	—	CLLR, BLR	Facebook: 48 terrorist users & 96 non-violent radicals	Radicals in Pakistan	I, A
11	Online Radicalisation Along a Continuum: From When Individuals Express Grievances to When They Transition into Extremism	Al-Saggaf	2018	Discusses radicalization models	RPAS	Not Specified	Facebook, Twitter, Youtube	Non-specific	I, P
12	Understanding the Radical Mind: Identifying Signals to Detect Extremist Content on Twitter	Nouh et al.	2019	Utilized psychological, emotional and personality theory	Tf-IDF, Word2Vec	RF, NN, SVM, KNN	Kaggle: Pro-ISIS, ”Normal” tweets, Kaggle: ISIS-content by non-radicals	Mainly ISIS	I
13	Social networks data analysis with semantics: application to the radicalization problem	Barhamgi et al.	2018	Ontology, Expert defined indicators	Preprocessing	Future work	Kaggle: Tweets Radical and neutral	Islamist extremism	P, I
14	Research on Online Supporting Community of Extreme Organization by AI-SNA Based Method	Xu et al.	2017	—	Word embedding, frequency	NB, LSTM	Twitter: 30,000 extreme + NLTK toolkit: 20,000 ordinary tweets	X terrorist organisation from Asia	I, A
15	Mining Pro-ISIS Radicalisation Signals from Social Media Users	Rowe and Saif	2016	Previous work related to the radicalization process.	Data-mining, BoW	None specific algorithm	Twitter: Pro-ISIS, anti-ISIS, neutral Seed set: O’Calleghan	ISIS in Europa	P
16	Supporting the identification and the assessment of suspicious users on Twitter social media	Tundis et al.	2018	—	BoW, DTF, n-gram	Future work	Twitter: 249 potential dangerous users	OC, TN	I, (P)
17	personality2vec: Enabling the Analysis of Behavioral Disorders in Social Networks	Beheshti et al.	2020	Golden standards for personality, behavior and attitude	personality2Vec, Word2Vec	CNN	17,000 tweets from users positive to extreme behavior	Non-specific	A
18	Defining an Alert Mechanism for Detecting likely threats to National Security	Cárdenas et al.	2018	(Emotions and behaviour)	Preprocessing	GBM, RF, DL	Twitter: Disruptive & non-disruptive	Non-specific	P, (I)
19	Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools	Gaikwad et al.	2021	—	TF-IDF with uni-gram	SVM, DT	Title, Abstract, Keywords from 64 studies	Extremism in general	P
20	Combating the challenges of social media hate speech in a polarized society: A Twitter ego lexalytics approach	Udanor and Anyanwu	2019	—	Preprocessing, n-gram, BoW	Unsupervised clissifier & future work	Nigerian tweets: negative or positive	Hate speech	I
21	Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning	Rehman et al.	2020	Want to include psychological dimensions in future work	Preprocessing, TF-IDF	NB, RF, SVM	Five distinct Twitter datasets	ISIS	I

Table 5: Chosen literature performing SLR.

ID	Title	Author	Year	Psychological theories	NLP Elements	Algorithm(s)	Data Set	Extremism	Predict, Identify
22	Predicting Online Extremism Content Adopters, and Interaction Reciprocity	Ferrara et al.	2016	—	—	LR, RF	Twitter: Suspended ISIS accounts, accounts exposed to ISIS	ISIS	P, I
23	Using KNN and SVM Based TF One-Class Classifier for Detecting Online Radicalization on Twitter	Agarwal and Sureka	2015	—	Linguistic features	KNN, SVM	Two existing sets combined with one self-conducted. (Only English)	Not specified	I
24	Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter	Benigni et al.	2017	—	—	RF, Iterative Vertex Clustering and Classificaion	Snowballed from 5 ISIS Twitter users	ISIS, online extremist community	A, I

Table 6: Recommended literature by the supervisor.

ID	Title	Author	Year	Psychological theories	NLP Elements	Algorithm(s)	Data Set	Extremism	Predict, Identify
25	Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter	Lara-Cabrera et al.	2017	Behavior indicators			3 Twitter datasets containg pro-ISIS and random users	ISIS	P

Table 7: Snowballed literature from review literature.

Reference	Title	Note
[Knudsen, 2020]	Measuring radicalisation: risk assessment conceptualisations and practice in England and Wales	It presents 22 risk indicators and how they are used in counter-terrorism.
[Skleparis and Knudsen, 2020]	Localising 'radicalisation': Risk assessment practices in Greece and the United Kingdom	Presents anti-radicalization policy in UK and Greece.
[Winter et al., 2021]	A moral education? British Values, colour-blindness, and preventing terrorism	Building resilience towards extremism at the school.

Table 8: Other recommended literature.

E Collected Dictionary

In the experiment in subsection 5.4.1, the weights of the terms used in the extreme corpus were calculated, and the ten terms considered to be the most important using TF-IDF are presented with a belonging weight value in section 6.2. The same experiment produced a dictionary of the 500 terms considered the most important ones to determine if a post is extreme using the adapted version of IDF. The order of the terms is based on how important the term is considered, where the earlier in the list the term is, the more important it is considered. It is not performed any manual operations to exclude terms but is solely created using an implementation of code.

trump, maga, like, democrats, kek, people, white, covid, trudeau, cdnpoli, gab, get, biden, one, twitter, good, time, new, hillary, memes, right, joe, joebiden, world, go, meme, want, know, women, would, altright, us, elxn, love, see, memewars, vaccines, make, justin, think, never, day, back, even, man, news, black, going, anti, chan, america, frens, first, still, two, let, mememagic, president, got, pol, really, say, real, antifa, msm, liberals, need, please, live, blm, great, fuck, also, today, take, stop, every, everyone, happy, whites, much, post, hate, breaking, made, life, anon, media, way, election, qanon, reddit, canada, years, year, old, lol, video, kun, vaccine, race, best, ever, winning, gabfam, fucking, usa, freespeech, shit, free, jews, big, many, draintheswamp, god, watch, must, better, nothing, thing, look, little, nick, censorship, immigration, follow, fakenews, well, anyone, prime, gay, left, may, minister, morning, coronavirus, could, pepe, country, keep, dnc, bad, says, always, hey, elliott, christmas, war, alan, stopthesteal, getting, clinton, men, use, obama, gregory, hollywood, justintrudeau, presidenttrump, american, th, guy, woman, help, tell, children, praisekek, cnn, something, support, speech, apu, literally, speakfreely, work, give, three, mememagicisreal, last, long, politics, sjws, gonna, israel, everything, groyper, sure, another, around, find, makes, said, yes, actually, show, vote, truth, stupid, based, someone, thanks, wants, imagine, state, family, fake, propaganda, trudeaumustgo, americans, call, come, money, brexit, looks, jewish, social, cozy, true, facebook, things, gets, wikileaks, freedom, anything, refugees, pegida, third, theleft, wwg, full, oh, racism, wga, feminism, share, read, trying, house, dumb, kill, potus, night, sex, racist, lot, put, kids, bill, name, person, without, hope, yet, police, hillaryclinton, canadians, four, wrong, pro, girls, china, account, government, trumpwon, open, sjw, rarepepe, faggot, next, banned, rd, believe, thank, art, funny, metoo, away, globalists, dindus, ass, remember, end, dead, change, maybe, power, rape, themedia, already, saying, fuentes, whitegenocide, communists, since, future, called, else, hard, votetrump, feminists, party, try, start, used, impeachment, ago, donald, home, political, face, hitler, wall, red, groypers, times, feel, socialjustice, aoc, enough, pelosi, youtube, *tonight, making, saveamerica, job, uk, democrat, care, run, coming, win, jew, globalism, blacks, fren, film, fact, socialists, liberal, death, wait, kylerittenhouse, evil, word, matter, fight, steal, fun, mask, yeah, britfam, nancy, internet, genocide, friends, pandemic, mean, europe, putin, group, nancypelosi, check, thought, gae, nice, hell, using, point, dems, lmfao, talking, jesus, place, debate, okay, means, seems, history, sleepyjoe, five, movies, found, health, done, nazis, become, comedy, earth, head, pretty, idea, gun, reality, needs, crazy, week, public, dr, million, child, bernie, far, watching, might, guns, words, ok, bitch, communism, climatechange, corruption, reason, trumplandslide, º°, part, bidenharris, science, guys, retarded, afd, die, soon, self, probably, americafirst, nazi, rittenhouse, calling, less, seen, stand, days, mind, human, talk, control, clownworld, kyle, soros, vaccinated, gender, nigga, unvaccinated, looking, young, andrew, followers, lives, borders, guess, gabriots, sauce, george, merrychristmas, though, eat, girl, law, taking, goes, gettr, six, least, enjoy, patriots, question, trust, canadian, st, fbi, instead, russia, stay, pm, lost, king, podestaemails, honkhonk, ask, cool, christ, email, crime, toronto, problem, knows, cannot, join, others, nation, gop, brandon, fresh, feeling, school, newright, lmao, non, communist, high*

F Word2Vec

This section delivers the 50 most frequent words in the extreme dataset. Each of the words is presented in Table 9 with the ten most similar words according to the first word2vec model described in subsection 5.4.2.

Word	Ten most similar words
Trump	maga, democrats, president, hillary, presidenttrump, potus, winning, usa, supporters, news
Trudeau	justin, cdnpoli, canada, liberals, canadians, justintrudeau, elxn, fakefeminist, teamtrudeau, trudeaumustgo
People	us, think, way, also, know, still, want, one, would, even
Like	know, get, also, even, think, make, say, people, better, still
Democrats	trump, dnc, maga, hillary, hillaryclinton, theleft, deepstate, imwither, obama, hollywood
Kek	praisekek, shadilay, memewars, mememagic, kekistan, dankmemes, dankmeme, theleft, childabuse, topkek
Covid	vaccines, pandemic, vaccine, coronavirus, vaccinated, woodfill, covidvaccines, billgates, please, jabs
Maga	trump, draintheswamp, presidenttrump, potus, winning, lockherup, crookedhillary, democrats, hillary, trumptrain
Cdnpoli	trudeau, justintrudeau, teamtrudeau, elxn, justin, trudeaumustgo, canada, liberals, fakefeminist, canadians
Get	like, getting, see, really, go, make, still, say, people, know
Gab	twitter, post, much, internet, follow, think, foolery, platform, magento, people
One	people, know, still, every, think, say, us, would, even, also
Twitter	internet, gab, post, meme, please, much, account, facebook, posts, liberals
White	black, whites, people, want, think, also, jewish, like, jews, really
Hillary	podestaemails, hillaryclinton, trump, clinton, maga, democrats, billclinton, weiner, imwithher, wikileaks
Time	day, years, way, life, think, one, still, year, many, world
Biden	president, one, us, still, put, trump, say, joewalsh, bff, johngritt
Good	like, best, know, great, love, even, well, bad, done, real
Justin	trudeau, cdnpoli, canadians, canada, justintrudeau, liberals, prime, elxn, fakefeminist, minister
Would	also, way, people, actually, could, never, want, one, say, going
Women	children, men, woman, young, goodadvise, rape, girls, trans, sex, feminazis
Memes	meme, mememagicisreal, memewars, mememagic, kek, shitlords, praisekek, shitposting, trump, dankmemes

Know	say, think, really, like, people, way, even, one, get, make
New	back, still, one, like, around, top, done, become, see, last
Us	people, one, never, want, everyone, everything, way, think, would, know
Want	think, people, way, need, would, also, going, know, like, get
Meme	memes, please, share, post, twitter, memiacs, gae, trump, democrats, memewars
Right	people, also, way, know, would, yet, like, want, well, one
Chan	pol, anon, kek, memewars, fakeraptorsfan, cdnpoli, trudeaumustgo, memefarmers, lpc, mememagic
Pol	chan, anon, fakeraptorsfan, thestormishere, geneticist, kek, lobstergate, bricktamland, torontostong, obamabiden
News	msm, trump, cnn, fakenews, maga, memes, media, president, presidenttrump, hillary
Canada	trudeau, canadians, cdnpoli, justintrudeau, liberals, justin, canadian, pmjt, ontario, elxn
Go	get, take, say, make, know, come, see, going, way, like
Think	know, people, say, way, want, one, like, much, never, even
Never	ever, say, would, think, people, us, one, want, still, everything
Memewars	mememagic, kek, mememagicisreal, dankmemes, praisekek, memes, shadilay, raussiagate, dankmeme, shitlords
See	get, make, know, really, like, keep, still, people, go, say
Joe	surrenderer, hunter, president, fmr, pstupid, arns, joebiden, pardons, poopypants, beatnik
Liberals	cdnpoli, trudeau, canada, justin, canadians, justintrudeau, trudeaumustgo, elxn, liberal, fakefeminist
Elxn	trudeaumustgo, cdnpoli, fakefeminist, trudeau, chooseforward, teamtrudeau, justintrudeau, canpoli, polcan, fucktrudeau
Even	still, know, people, like, way, something, think, actually, one, literally
Make	get, know, still, see, really, like, way, give, makes, also
Day	time, year, today, one, years, life, week, place, hope, way
Man	people, never, woman, would, davidcopafeel, one, guy, also, still, years
Hate	disavowhillary, say, jews, people, literally, crime, support, jocks, even, oldtestament
Going	want, people, may, would, way, could, still, already, get, many
Back	still, away, well, one, like, never, right, friends, better, done
Love	good, great, like, help, best, maybe, one, think, together, need
President	trump, election, biden, votes, years, joe, donald, vote, democrats, americans
First	also, every, better, great, back, one, like, different, americanhistory, hypergamous

Table 9: The most similar words to the top 50 used words in the extreme corpus using model one.

Ingrid Vrålstad Løvås

Recognizing Social Media Right-Wing Radicalization

NTNU

Kunnskap for en bedre verden