Master's thesis

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

Thomas Alejandro Fernandez Ramirez

How to Catch a Far-Right Radical

A model to detect far-right/alt-right radicalisation of social media users at individual and group level

Master's thesis in Master of Science in Informatics
Supervisor: Björn Gambäck
July 2022

NTNU
Norwegian University of
Science and Technology

Thomas Alejandro Fernandez Ramirez

How to Catch a Far-Right Radical

A model to detect far-right/alt-right radicalisation of social media users at individual and group level

Master's thesis in Master of Science in Informatics
Supervisor: Björn Gambäck
July 2022

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

NTNU

Kunnskap for en bedre verden

Thomas Alejandro Fernandez Ramirez

How to Catch a Far-Right Radical

A model to detect far-right/alt-right radicalisation of social media users at individual and group level

Master Thesis, Spring 2022

Artificial Intelligence Group
Department of Computer and Information Science
Faculty of Information Technology and Electrical Engineering

Abstract

Social media platforms are becoming the new arena for far-right extremists. The extremists communicate, plan attacks, and radicalise new users through the internet. With the increasing growth of users joining these platforms, the task of detecting far-right radical users has become challenging. Far-right radicalisation is now digital and more dangerous than ever. The goal of this Master's thesis is to explore detection methods from other types of radicalisation and try to adapt and tailor these techniques for far-right radicalisation detection.

The goal was broken down into tasks and research questions. The first research question focuses on finding a potential method for the detection of far-right radicalisation. This method is the primary method in the study. The second research question focuses on creating datasets containing far-right and regular users. Regular users refer to the non-radical users of a platform. The third research question focuses on creating a method for extracting radical terms relevant to far-right users, and thereby creating a radical dictionary. The fourth research question investigates potential improvements with three suggested changes. The first modification adds more radical terms used in the method, the second changes the way radicalisation is evaluated, and the third adds two new metrics: profanities and average post length.

The selected method stems from the detection of Islamic radicalisation. It was chosen due to its excellent performance (F1=0.901) when trained on classifiers, and because it was based on a social science theory which distinguishes it from other methods. The method calculated radicalisation on three different levels: micro (individual), meso (groups), and macro (society). Macro was excluded due to the complexity of extracting textual data from multiple sites. The radical dataset was constructed from a far-right social media site called Gab and gave 291 users with 75 788 posts. The regular dataset was constructed from Twitter and contained 213 users with 56 299 posts. An experimental method retrieving radical terms was created by using far-right manifestos with the keyword extractor, KeyBERT. The method returned 2 764 terms.

The tailored implementation in the first research question with the datasets and radical dictionary showed poor results with an average F1 score of 0.569. The test was performed twice, in which micro and meso were tested separately. The average F1 score is the result of both these tests. The different modifications in the fourth research question gave varying results. The first modification got even worse results, with an average F1 score of 0.520. The second modification slightly improved the results, with an average F1 score of 0.659. The third modi-

fication returned substantially improved results with an average F1 score of 0.857.

When combining all the features and not separating the values of radicalisation on the individual and the group lever, the results changed drastically. The best score combined two profanity metrics and two average lengths of posts with micro and meso. The result achieved was 0.947 in F1.The results are encouraging, but detection of far-right radicalisation needs further research.

Sammendrag

Sosiale medier er på vei til å bli den nye arenaen for høyreekstreme. Ekstremistene kommuniserer, planlegger angrep og radikaliserer nye brukere via internett. Etter hvert som antallet brukere som blir med på disse plattformene øker, har det blitt en utfordring å oppdage høyreekstreme radikale brukere. Radikalisering av høyreekstreme brukere er nå digitalt og farligere enn noen gang. Målet med denne masteroppgaven er å utforske oppdagelsesmetoder fra andre typer radikalisering og forsøke å tilpasse og skreddersy disse metodene for å oppdage høyreekstreme radikaliseringer.

Hovedmålet med oppgaven har vært delt inn i oppgaver og forskningsspørsmål. Det første forskningsspørsmålet fokuserer på å finne en potensiell metode for å oppdage høyreekstremistisk radikalisering. Denne metoden er den primære metoden i studien. Det andre forskningsspørsmålet fokuserer på å lage datasett som inneholder høyreekstreme og vanlige brukere. Med vanlige brukere menes de ikke-radikale brukerne av en plattform. Det tredje forskningsspørsmålet fokuserer på å lage en metode for å trekke ut radikale termer som er relevante for høyreekstreme brukere, og dermed lage en radikal ordbok. Det fjerde forskningsspørsmålet undersøker potensielle forbedringer og dermed introduseres tre foreslåtte endringer. Den første modifikasjonen endret antall radikale termer inkludert i metoden, mens den andre endret måten å kalkulere radikalisering. Den tredje modifikasjonen la til to nye målte verdier: banneord og gjennomsnittlig lengde på poster.

Den valgte metoden har sitt utspring i oppdagelsen av islamsk radikalisering. Den ble valgt på grunn av sin utmerkede ytelse (F1 = 0,901) når den ble trent på klassifiserere, og fordi den er basert på en samfunnsvitenskapelig teori som skiller den fra andre metoder. Metoden beregner radikalisering på tre ulike nivåer: mikro (individuelt), meso (grupper) og makro (samfunn). Makro ble forkastet fordi det var for utfordrende å trekke ut tekstdata fra flere steder. Det radikale datasettet ble konstruert fra et ekstremt høyre sosialt medienettsted kalt Gab og genererte 291 brukere med 75 788 innlegg. Det vanlige datasettet ble konstruert fra Twitter og inneholdt 213 brukere med 56 299 innlegg. En eksperimentell metode for å finne radikale termer ble laget ved hjelp av høyreekstreme manifester med nøkkelordsekstraktoren KeyBERT. Metoden ga 2764 termer.

Den skreddersydde implementeringen av det første forskningsspørsmålet med datasettene og den radikale ordboken viste dårlige resultater med en gjennomsnittlig F1-score på 0,569. Testen ble utført to ganger, hvor mikro og meso ble testet hver for seg. Gjennomsnittlig F1-poengsum er resultatet av begge disse testene. De ulike modifikasjonene i det fjerde forskningsspørsmålet ga varierende

resultater. Den første modifikasjonen ga enda dårligere resultater, med en gjennomsnittlig F1-score på 0,520. Den andre modifikasjonen var bedre, med en gjennomsnittlig F1-score på 0,659. Den tredje modifikasjonen ga bedre resultater med en gjennomsnittlig F1-score på 0,857.

Når man kombinerer alle verdiene og ikke skiller verdiene av radikalisering på individet og gruppe-nivå, endret resultatene seg drastisk. Den beste poengsummen kombinerte to banningsmålinger og to gjennomsnittlige lengder på innlegg sammen med mikro og meso. Resultatet som ble var 0,947 i F1. Resultatene er oppmuntrende, men ytterligere forskning er nødvendig for å oppdage radikalisering av høyreekstreme grupper.

Preface

This Master's thesis is the final part of achieving the Master of Science in Informatics specialising in Artificial intelligence from the Norwegian University of Science and Technology in Trondheim, Norway. I want to thank my supervisor Björn Gambäck, who helped me with the thesis and supported me during challenging times. Thanks for the excellent conversations and guidance. I would also like to show my gratitude to my lovely girlfriend, Synne Sandberg, who has been my biggest supporter in periods of doubt and my anchor throughout this whole process. Also, I thank her for her patience when proof-reading this thesis. A big thanks to my friend Sander Lindberg for helping me in a time of need.

Thomas Alejandro Fernandez Ramirez
Trondheim, July 7, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the world, the digital revolution has begun. The human population is at 7.91 billion by January of 2022, and the numbers indicate it will reach eight billion by the middle of 2023. At the beginning of 2022, the number of global Internet users was 4.95 billion, which means that 62.5% of humanity is currently on the web. Furthermore, the current number of social media users globally is 4.62 billion, which means that almost 58.4% of the current human population are on social media platforms [Kemp, 2022]. The world is getting more interconnected, and rapid communication nowadays reaches all the world's corners. The internet has provided humanity with one of the most revolutionary inventions where people can share thoughts and beliefs with others. For the most part, this has positive ripple effects. However, there are people who exploit the internet's availability to spread extremist beliefs to easy prey.

1.1 Motivation and Background

The motivation for this Master's Thesis is to develop a system that detects and finds *far-right* people who are currently becoming radicalised on social media platforms. Social media platforms allow users to voice their ideas and opinions, and provide a tremendous amount of information to explore. The thesis aims to explore the findings from research on other radicalisation types and implement them in order to detect far-right/alt-right users. The results will be analysed and compared to the current state-of-the-art in detection of radicalisation. Furthermore, modifications and adjustments will be explored to achieve even better results. Four key points are fundamental for the motivation and will be presented in the following sections.

1.1.1 A New Threat with Old Roots

As of 2022, the threat of politically motivated violence has increased in Norway. In the National Threat Assessment (NTA) of 2022, the topic of *Politically motivated violence extremism* was included. The NTA is an annual report published in cooperation between Norwegian Police Security Service[1], Norwegian Intelligence Service[2] and National Security Authority[3], and aims to give the public a comprehensible overview of the current threat to the Norwegian Kingdom. The threat level for general terror is categorised as *moderate*. However, the execution of a terrorist act by *right-wing extremists* is categorised as *even chance*, which means that it is 40-60% likely to happen in 2022. Arguments supporting the beliefs of increased danger include the rise of conspiracy theories surrounding the government's handling of the CoVid-19 pandemic and increased digitalisation [Gjørv et al., 2021].

Suggested in Jenkins [2022], far-right extremists have for a long time been organised in a leaderless hierarchy and individually carried out attacks on behalf of a more significant cause. However, modern domestic extremists have begun to coordinate through social media platforms. The new far-right extremists are now online.

1.1.2 From Jihad to Far-right

The growth of ISIS presented a new form of online radicalisation. On June 9, 2014, ISIS declared the plan to establish an Islamic caliphate in the Middle East. In contrast to earlier terrorist organisations such as Al Qaeda, Al-Shabaab, Hamas, and Hizbollah, the internet and mainstream media were used to promote ISIS. The propaganda contained brutal executions, destruction of monuments and artwork, and murder [Andersen and Sandberg, 2020, p.1506-1507]. According to a study from January 2015, 20,700 foreign fighters joined the conflict in Syria and Iraq from 2011-2015 [Neumann, 2017, 85-87]. ISIS caught the world's attention when young people from around the world joined the fight, supporting the Islamic state. The surge of ISIS and its latest form of radicalisation created a demand for research on online radicalisation.

The research shows multiple promising methods to detect radicalisation. However, these methods have currently only been used on Islamic radicalisation. The motivation is to observe whether similar results can be demonstrated by using the same approaches on right-wing radicalisation.

[1]Norwegian name: Politiets sikkerhetstjeneste(PST)
[2]Norwegian name: Etterretningstjenesten(E-tjenesten)
[3]Norwegian name: Nasjonal sikkerhetsmyndighet

1.1.3 From Twitter to Gab

Empirical evidence based on the findings in Ramirez [2021] suggest that Twitter is the preferable data source for research. Twitter delivers a high-standard API service, making it easy and safe to extract data. The API service allows the search of terms, hashtags, people, and posts. Twitter also wants to prevent behaviour that discourages users from interacting with the site and focuses on moderating content that is in conflict their terms and conditions [4]. In 2017, in response to the rally in Charlottesville, Twitter started to actively remove far-right activists and far-right organisations from the platform. This created a demand for social media platforms with less moderation. The ecosystem of alternative platforms, called Alt-Tech, emerged as a solution for banned far-right users. The platform Gab is a social networking platform similar to Twitter. Gab has since then become a popular platform for far-right extremists and has claimed to be growing by 10,000 users each hour [Jasser et al., 2021, p.2].

Due to the minimal research conducted on Gab as a place for far-right extremists, there is a desire to study this potential new data source. The resemblances between the platforms assure that the method implemented on Twitter-data is compatible with data extracted from Gab. Since the site hosts multiple far-right users, it is the perfect source for creating a dataset containing radical users.

1.1.4 Preventive Work Against Extremism

After decades of the war on terrorism, results have shown that the task of neutralising terrorists and sabotaging planned terror attacks is not enough [Borum, 2011, p.8]. The mission should concentrate on detecting and preventing radicalisation at an earlier stage and, at the same time, preventing terrorist attacks. The task is not *what* people believe but rather *how* they end up believing as they do. As said in Borum [2011], a successful framework for the detection requires mechanisms utilising the micro (individual) and the macro (cultural) levels of radicalisation. The framework will also require adjustments for each type of radicalisation, making it less likely that a framework will be "one-size-fits-all" [Borum, 2011, p.8].

Radicalisation is the earliest indication of a person adopting a more extremist ideology. This dangerous process must be stopped as early as possible. To prevent future terrorist attacks, it is beneficial to view radicalisation as a point of attack in the work against terrorism.

[4] {https://help.twitter.com/en/rules-and-policies/twitter-rules}[Accessedon08.07.2022]

1.2 Preliminary Study

As a part of the Master's degree in Informatics at NTNU, the penultimate semester is used to attain knowledge on the topic of choice for the Master's thesis. The goal of this preliminary study [Ramirez, 2021] is to explore the domain. The aim is to gain fundamental knowledge about the domain and have enough ability to develop a research goal for the Master's thesis. The preliminary studies for this Master's thesis were written in Autumn 2021, where the research goals and research questions were:

Goal *To acquire an overview of the field by predicting users vulnerable to radicalisation on social media*

RQ1 *What methods exist for prediction of radicalisation traits?*

RQ2 *What methods are used to find datasets of radical content?*

The "goal" of the preliminary study overlaps with the topic of this Master's thesis. Thus, this thesis can be seen as a continuation of the preliminary study. The theoretical part and structural literature review were particularly relevant for this Master's thesis, which is why they are included, as seen in Chapter 2 and Chapter 3. Chapter 2 is based on the *Background Theory* chapter from the preliminary study and demonstrates theories from computer science and other science fields. The theories included are based on the papers retrieved from the literature search. Chapter 3 contains a Structured Literature Review (SLR) that was performed in the preliminary study. The goal was to methodically find research papers covering the topic of radicalisation. The SLR process from the preliminary study is used because the results yielded the necessary foundation for this Master's thesis.

1.3 Definitions

The following section presents important definitions of terms that play a crucial role in the study of radicalisation. The choice to create a list of definitions is rooted in the fact that various terms can be seen as "new" and have recently become a part of the daily speech. Moreover, many terms are often used synonymously, which can cause confusion. Finally, some terms do not have a universally accepted definition, e.g. terms such as far-right and alt-right. The purpose of presenting the definitions in this section is to secure that the terms are clear throughout the thesis.

1.3.1 Radicalisation

Radicalisation is a complex term with multiple definitions. The term does not have a universally accepted definition and differs from research to research. In discussion surrounding the term's definition Borum [2011] argues that it can be split into two types of definitions. The first definition focuses on violent radicalisation, which centres around the active pursuit or acceptance of using violence to achieve the stated goal. The second definition introduces a broader sense of radicalisation that emphasises the active pursuit or acceptance of far-reaching changes in society. This definition also states that radicalisation may or may not involve the threat or use of violence to achieve the stated goal and may or may not pose a danger to democracy [Borum, 2011, p.12-13].

The definition used throughout the thesis is more associated with the second definition, *where a person is in the process of accepting a belief that constitutes violence as a necessary means to attain the stated goal.* Given that it is a gradual change in an individual's beliefs, radicalisation is assumed to *be a process.* It is important to note that a person who is in the process of becoming radicalised is also referred to as "a radical".

1.3.2 Extremism

Where radicalisation ends, extremism begins. Extremism is also divided into two different definitions, in which one is used to refer to ideology while the other is about the individual's accepted methodology for acquiring the desired goal. The ideological aspect defines extremism as contrary to the core values in society, while the methodological aspect is when an individual believes that violence is an acceptable means to achieve the stated goal [Borum, 2011, p.10]. Note that an extremist has accepted violence as a means to attain a goal, while a radical is in the process of accepting the same view on violence.

Following that notion, the definition used throughout the Master's thesis is associated with the second definition, *where a person has already accepted a belief that violence is seen as a reasonable means to acquire the desired goal.* An extremist is therefore assumed to be a person that has completed the radicalisation process. Thus, due to their active use of violence, people who have executed terrorist attacks are considered extremists.

1.3.3 Far-right

Far-right, also known as right-wing extremism, refers to the right side of the political spectrum (see Figure 1.1) and does not have a generally accepted definition.

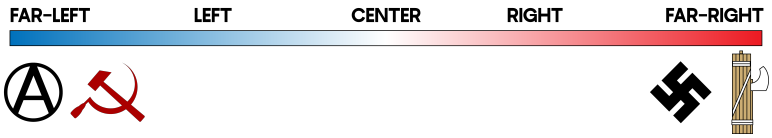

Figure 1.1: The political spectrum. Symbols from left to right: Anarchism, Communism, Nazism and Fascism

One reason it has not received an accepted definition is partly because the term is not only used for scientific purposes but also for political purposes [Mudde, 2000, p.10]. Most researchers that have tried to define the term have concluded that far-right is an ideology containing borrowed elements from other ideologies. Based on the conclusion in Mudde [2000], five features were mentioned the most times throughout the 26 different definitions. These features are nationalism, racism, xenophobia, anti-democracy and the strong state [Mudde, 2000, p.10-11].

Subgroups found in the right-wing/far-right part of the spectrum are Fascism and Nazism [Mudde, 2000, p.11-16]. "Newer" versions of Fascism and Nazism are referred to as Neo-Fascism and Neo-Nazism. They are based on the desire to restore the foundation of the original ideologies.[5]

1.3.4 Alt-right

The word alt-right is an abbreviation and means *Alternative right*. The term was introduced by Richard Spencer and referred to the resurgence of a new right-wing movement in early 2008 [Hawley, 2018, p.51-53]. The term initially had a short lifespan and died out before re-appearing in 2015 with greater force on sites such as 4chan and Reddit. Meme culture was shaped around the term and a meme known as "Pepe" took the lead as the digital mascot [Hawley, 2018, p.70]. Pepe, also known as "Pepe the Frog", is a humanoid-amphibian figure that went viral in 2010 and was crowned the "biggest meme of the year" in 2015. The meme was later classified by the U.S. Anti-Defamation League as a hate symbol [Glitsos and Hall, 2019].
The Alternative-right attempts to distance itself from conservatism (specifically American conservatism) [Hawley, 2018, p.92], due to it having more in common with skinheads, the KKK and National Alliance [Hawley, 2018, p.70-71] concerning race [Hawley, 2018, p.11], and particularly white supremacy. The alt-right shares the ideology of white supremacy and anti-immigration policies but does not share the harsh rhetoric of older Conservative movements. Instead, alt-right

[5]The prefix *neo* stem from Ancient Greek and means "New, young".

views are delivered using an ironic approach on the internet [Hawley, 2018, p.71]. Contrary to conservatism, religious beliefs are not necessarily connected to the ideology, and, in some cases, the alt-right criticises Christianity [Hawley, 2018, p.100]. Similarities to the rhetorics found in Nazism can be found in the "meme" culture online by alt-right users. The alt-right has been observed where they "secretly" single out Jewish people on the internet by putting three parentheses by their name [Hawley, 2018, p.83] or referring to them as "Globalist" [Hawley, 2018, p.124]. The term alt-right holds various values and elements from other "-isms", which makes it challenging to create a clear definition. Nevertheless, to summarise the alt-right, this comment given by a person from the *Daily Stormer* explains the alt-right as:

> The core concept of the [alt-right], upon which all else is based, is that Whites are undergoing an extermination, via mass immigration into White countries which was enabled by a corrosive liberal ideology of White self-hatred, and that the Jewish elites are at the center of this agenda, even Milo himself admitted this [Hawley, 2018, p.141].

In this Master's thesis, the definition of the *alt-right* is based on far-right ideologies where race, specifically white supremacy, is an important matter. An alt-right person can also be defined as far-right since the alt-right borrows features from other ideologies on the right end of the political spectrum.

1.4 Goals and Research Questions

The ensuing goal and research questions for this Master's thesis are based on the motivation and background presented in Section 1.1. The research questions are written in abbreviations where "RQ" and their number are combined. For instance, research question one is RQ1.

Goal *Detecting political radicalisation of users on social media.*

The thesis will focus on far-right and alt-right radicalisation. There is extensive research on radicalisation, but not as much on far-right and alt-right radicalisation. The goal is to select a model that has been used to track a different type of radicalisation and adjust it to detect right-wing radicalisation. In order to attain this goal, it was essential to create a far-right dataset and an ordinary dataset, as well as a list of far-right terms. The dataset with far-right users will contain far-right users on social media, while the ordinary dataset contains regular users. The model, datasets, and list of terms are used to investigate the performance of a few selected machine learning classifiers. Furthermore, significant attention will be given to improving the model.

Research question 1 *Is there a method for detecting far-right users within another domain of radicalisation detection?*

The research question aims to explore the established methods developed in other fields of radicalisation. The aim is to find a suitable model for detecting far-right and alt-right users on social media platforms with few modifications.

Research question 2 *Does a dataset suitable for detecting far-right users exist? If not, how can it be created?*

The research question strives to find datasets containing far-right or alt-right users suitable for the models selected in RQ1. If there are no relevant datasets, an approach to construct the datasets will be investigated and executed. The dataset will also be used on the modified model in RQ4.

Research question 3 *What are the advantages and disadvantages of using an unsupervised method to find radical terms?*

The research question aims to find a method to create far-right and alt-right terms. The experimental approach will be analysed based on the returned terms. The radical terms found here are used throughout the Master's thesis. The model will take in use unsupervised technologies to discover terms automatically..

Research question 4 *How can the existing approach for detecting radicalisation be improved?*

The research question focuses on improvement areas for the implementation of the model from RQ1. The results from RQ1, RQ2, and RQ3 will be analysed to suggest possible modifications. The modifications will be implemented and compared to the results in RQ1.

1.5 Research Method

Various methods were used to fulfill the goals and answer the research questions for this thesis. Research question 1, called RQ1, required an overview of the current state-of-the-art on detection of radicalisation. A structured literature review (SLR) inspired by Kofod-Petersen [2015] was used to find related works in the field. The execution of the SLR is described in Chapter 3 together with the discoveries from the related work. The discoveries are the foundation of the research for this thesis and, according to the findings, only one model was suitable. The model was implemented with the datasets from RQ2 and a list of radical terms (radical dictionary) from RQ3 to analyse its performance in detecting far-right and alt-right users on social media. Finally, the model was

implemented on three different machine learning classifiers. Research question 2 uses the research and findings from Chapter 3 to find the adequate dataset for the model. No suitable datasets were found in the research, which is why a new dataset was created. The design of the approach to create a new dataset was based on the methods found in related works and later performed. The results were analysed and evaluated through the implementation of RQ1 and RQ4. Research question 3 aims to create a method that automatically generates far-right radical terms. The created method has advantages and disadvantages and these are discussed and compared to methods made by domain experts. The radical terms produced by the new method were evaluated based on the results in RQ1 and RQ4 and examined manually. Research question 4, RQ4, created three different suggestive modifications to the model in Research question 1. The modifications were based on research that suggested important features and parameter tweaking. The results achieved in RQ1 were used to compare the performance of the modifications. An analysis of the different results was compared to similar findings in other research. The modified models, similar to RQ1, were implemented on three different machine learning classifiers.

1.6 Contributions

- An alternative viewpoint on the task of detecting far-right users online.
- An overview of the current state-of-the-art research in the detection of radicalisation.
- A method of creating a far-right dataset by using Gab.
- An unsupervised method of creating far-right terms from manifestos by extremists.
- A suitable method for detecting far-right radicals with features based on research and academic studies.

1.7 Thesis Structure

Chapter 1 introduces the motivation and background behind the Master's thesis. Important definitions are defined and presented.

Chapter 2 presents the background theory and necessary information to understand the content of this Master's Thesis.

Chapter 3 presents the execution of a Structured Literature Review and literature relevant to the detection of radicalisation. The interesting findings are summarised and presented also. The literature presented here is the foundation for the Master's Thesis.

Chapter 4 presents the planning and execution of experiments. The experiments are designed to answer the research questions.

Chapter 5 contains the results from all the experiments from Chapter 4.

Chapter 6 contains discussions of each research question with the results from Chapter 5. The results will also be compared to similar findings from other research.

Chapter 7 concludes the Master's thesis and presents the limitations, contributions, and future work.

Chapter 2

Background Theory

This chapter will present relevant theories to understand the topics presented in this Master's thesis. The Section 2.1 presents the different social media platforms presented throughout the thesis. Section 2.2 will present necessary domain-specific theory and methods from Natural Language Processing. Continuing, Section 2.3, presents the field of Machine Learning and the theory behind different approaches, and explains how similarity/distance in Mathematics is used as a similarity measurement in ML. The last section of 2.4 presents the various methods used to evaluate performance in ML and what the metrics convey.

Sections 2.1, 2.2, 2.3, and 2.4 are from the Background Theory chapter in Ramirez [2021] and are included since it contains theory that still are up-to-date. There are minor changes to the sections to correct grammatical errors and structural errors, and does not change the content. Section 2.3.7, Section 2.5, Section 2.1.5, and Section 2.4.4 are new and added into the chapter.

2.1 Social Media

This section stem from the preliminary study and contains minor changes as mentioned in the introduction of this chapter. The original section can be found in [Ramirez, 2021, p.7-9]. Newly included section can be found in Section 2.1.5.

2.1.1 Twitter

Twitter is a social media launched in 2006. Users of the platform use the social media platform to publish micro-posts, referred to as a tweet. The tweets are of fixed character length limit; no post can surpass this limit. Initially, the

limit of the "tweets" was set to 140 characters but was later in 2017 doubled to 280 characters[1]. Each user registers with a unique username when joining the platform. User profiles contain profile images and a small biographical section where users can write about themselves. After registering, a user can start to interact with other Twitter users. Users can both follow or be followed by other users. By following a user, if the followed user publishes a tweet, the tweet will appear on their homepage. The homepage is their feed of the newest tweets by the users they follow. Users can also share other users' tweets, and this action is called retweeting. When creating a tweet, a user can label the content tags with tags. These tags are known as hashtags and have the #-symbol as a prefix. The user can select hashtags already made or create an entirely new hashtag. Hashtags can be used to interact with larger groups or social movements, such as #Eurovision and #BLM. Twitter is currently at rank #35 in the Alexa ranking[2] and one of the most used sites on the internet.

2.1.2 Facebook

Facebook was invented by Mark Zuckerberg together with his dorm friend at Harvard University [Hall, 2021]. Today, it is a multi-million enterprise far-reaching into each corner of the world. The social media platform has coverage in of the population of 35.6% [3] of the world's population in 2021. The service is free to use and makes it possible to interact with other users. Users create a profile with the possibility of uploading images, joining or creating groups, updating status, and more. Each profile page has a timeline showing changes and activities by friends. Friends on Facebook are users one as a user has accepted to be their "friends". One as a user can ask another user to become friends or be asked by them. One has to accept or deny the request. The service also provides interactions such as commenting and liking publications on the site. A chat function called Messenger makes it also possible to communicate in real-time with the friends [Hall, 2021]. In 2021, Facebook changed its name to Meta as they now shifted their attention to the future product "The Metaverse"[4]. The product Facebook is still called Facebook but is now organised as one of the products from Meta.

[1] `https://www.bbc.com/news/technology-41900880` [Accessed in 10.06.2022]

[2] `https://www.alexa.com/siteinfo/twitter.com` [Accessed in 14.12.2021]

[3] `https://www.statista.com/statistics/241552/share-of-global-\population-using-facebook-by-region/` [Accessed 27.11.2021]

[4] `https://about.fb.com/news/2021/10/facebook-company-is-now-meta/` [Accessed in 10.06.2022]

2.1.3 Gab

Gab is an alternative social media platform where "free speech" is a priority. Their primary mission is to "defend, protect and preserve free speech online for all people", as the site explains. Gab is similar to Facebook and Twitter, where posts referred to as "gabs" are limited to 300 characters. However, contrary to other social media platforms, moderation of publication is minimal. Gab state they only intervene when if the content is about illegal activity, threats of violence, doxxing[5], pornography, child exploitation, or spam [Goodwin, 2021]. It has been observed by online alt-right and far-right users the use of numbers and code to hide hate messages. Numbers as 88[6] or usage of triple parentheses around named. The number means the eighth letter in the alphabet, "H", referring to "HH". The "HH" refers to "Heil Hitler". The parentheses phenomena around a name to mark the person for having Jewish heritage [Tuters and Hagen, 2020].

As of April 2020, Gab reported 3.7 million monthly active users. The controversy around the site's non-moderation of content has created a hosting service for hate-filled posts, violent speech, misinformation spreading, and conspiratorial discussion room.[7]

2.1.4 Reddit

Reddit refers to itself as "the front page of the internet". The site contains millions of sub-communities called "subreddits". Each subreddit is a community that represents a topic. People join the subreddit and discuss with other community members interested in the same topic. The structure of the subreddits is like a forum, where members can publish posts available for anyone to see. The posts are voted by users and control their relevance to the community. Subreddits have the recognisable prefix "r/" to the community. The subreddit also has voluntary moderators who govern the subreddit's content and have the potential to remove a post or ban users from publishing. Reddit is also known for using a lot of Reddit jargon such as OP (Original Post), TIL (Today I learned), and AMA (Ask me anything)[8].

[5] Doxxing means publishing private information on the internet with malicious intentions

[6] `https://www.adl.org/resources/hate-symbol/88`[Accessed in 07.07.22]

[7] `https://edition.cnn.com/2021/01/17/tech/what-is-gab-explainer/index.html`[Accessed in 10.06.2022]

[8] `https://www.digitaltrends.com/web/what-is-reddit/`[Accessed in 10.06.2022]

2.1.5 90-9-1 Principe of Internet

The 1% rule, or 90-9-1 principle, where in internet communities, 90% of users will not contribute, while the 9% occasionally contribute, while the last 1% creates the majority of content. In the study, Trevor [2014] the conclusion shows this rule of thumb is consistent in four different sites and even suggests that the 1% is far less than 1%. The radical users, because of this bias, can be retrieved from the 9% or the 1% users.

2.2 Natural Language Processing

This section stem from the preliminary study and contains minor changes as mentioned in the introduction of this chapter. The original section can be found in [Ramirez, 2021, p.10-12].

Glossaries in NLP

The thesis will mention terms such as document, corpus, and vocabulary multiple times. When used in NLP, document a predetermined unit of text. It can either mean a book, a page, a sentence, etc. A corpus is a collection of documents. Vocabulary refers to all the unique words in a corpus, meaning all the words found inside the documents.

2.2.1 Representation of Text

In ML, algorithms can not use text directly and need the text to be converted to numbers. This needed step is referred to as preprocessing in NLP. Text is usually converted to vectors or tokens. This subsection will present the different techniques used in the field of NLP.

One Hot encoding

One Hot encoding (OHE) is a simple approach to representing a document. A vector is created for the document where each position has one unique word. If the word occurs in a document, it is represented in the vector as one, and zero if not. If the word appears multiple times, it will still get only the value one. The words used in the vectors can either be by a predetermined set of words or an entire vocabulary. The challenge with this approach is that text loses its order. For instance, the sentences "Mom loves Dad" and "Dad loves Mom" would be represented identically.

Bag-of-Words

The text is represented in a vector with the length of unique words, also known as the corpus vocabulary. A bag-of-words representation, however, differs from OHE as the vectors contain the frequency of words. The values in the vector represent the number of occurrences in the document. Similar to OHE, the challenges of this approach are that the order of the words is lost in the vector representation.

N-gram

Representation of text by N-grams is to take into consideration the neighboring sequences of tokens in the document. N refers to the size of tokens taken into consideration, called the window. An example can be seen in Figure 2.1 where unigram, bigram, and trigram. There N values are 1, 2, and 3, respectively.

```
   text = "it is raining today"
unigram = ["it", "is", "raining", "today"]
 bigram = ["it is", "is raining", "raining today"]
trigram = ["it is raining", "is raining today"]
```

Figure 2.1: Illustrations of Unigram, Bigram, and Trigram

2.2.2 Word Embedding

Different methods are used to try to inherit the semantic meaning of a word. Generally, methods covert the words to a position in a latent space. Distance in this vector space will represent similarity. For instance, the words king and queen are similar to each other and will be represented as similar vectors.

TF-IDF

TF-IDF is a method of weighting words in natural language processing and consist of Term Frequency and the Inverted document frequency. Term Frequency (TF) represents the frequency of a word in a given document. The Inverse Document Frequency (IDF) represents the word distribution in a corpus. IDF was introduced in 1972 by Karen Spärck Jones [Jones, 2004]. Term frequency counts the appearance of a term in a given document to represent the documents, while the inverted document frequency represents the importance of the word concerning the entire corpus. The idea is that if a word exists in multiple documents throughout a corpus, the term will most likely be a vague one and of low importance.

$$\mathrm{TF-IDF}(t, d, D) = tf_{t,d} \cdot \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2.1}$$

TF-IDF rewards unique terms with lower occurrence throughout a corpus with higher weights and punishes vice versa. The result is a vectorised representation of a document. It is calculated through Equation 2.1, where N represents the size of the corpus and D the given document. The $tf_{t,d}$ represents frequency of the term in a given document.

Word2vec

Word2vec is algorithm that uses neural networks to understand words relationships. The process was introduced in 2013 by Google in [Mikolov et al., 2013], and had two different methods; Continues Bag-of-word (CBOW) and Skip-Grams (SG). When **CBOW** is presented with a list of words, the algorithm will utilise the surrounding words, called context, to predict a word. **SG** tries to guess the context based on a single word, almost the reversed task preformed by CBOW.

GloVe

GloVe stands for **Glo**bal **Ve**ctors for Word Representation[9] and an unsupervised learning algorithm used to represent words. The approach is presented in Pennington et al. [2014], and uses the statistics of word-to-word occurrences from the entire corpus. Using an entire corpus, it gets the relation of the word by using all documents, hence the name global. The word's semantic meaning is derived from a co-occurrence matrix that tells how frequently a word pair appears together. This approach by using statistics gives each word in a unique vector representation, making it possible to find similar words, analogies, and more [Pennington et al., 2014]. In Pennington et al. [2014], GloVe was compared to other embeddings, such as word2Vec, and outperformed other models on word analogy, word similarity, and named entity recognition tasks.

2.3 Machine Learning

This section stems mostly from the preliminary study and contains minor changes as mentioned in the introduction of this chapter. The original section can be found in [Ramirez, 2021, p.12-18]. Section 2.3.7 is a new subsection added presenting the Transformers architecture and pre-trained Transformer models, such as BERT.

Machine Learning is a multidisciplinary field with the purpose of finding patterns in complex data. It contains elements from different fields, such as statistics

[9] `https://nlp.stanford.edu/projects/glove` [Accessed on 10.06.2022]

and probability, AI, computational complexity, control theory, information theory, philosophy, psychology, neurobiology, and more [Mitchell, 1997, p.3]. The definition of learning in Machine Learning mainly focuses on learning through experience on a specific task to improve. As seen in the suggested description of the Well-Posed Learning Problem in Tom Michell's book "Machine Learning" [Mitchell, 1997, p.3]:

> **Definition**: A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E

Machine Learning can be divided into three parts based on the available data during training, **Supervised** and **Unsupervised** learning [Goodfellow, 2016, p.104], and **Reinforcement Learning** [Goodfellow, 2016, p.106].

Supervised Learning uses a set of data together with the corresponding target label during training. In other words, the algorithm try to associate vector $\mathbf{x}$ with targeting vector $\mathbf{y}$; hence the name supervised. This is because it gets to know the correct label. They are mainly used in regression and classification problems. For instance, let us say you have three folders with images. The first is a folder with cat images, while the second folder contains dog images. The third contains cat and dog images in the same folder without any labels. By training a supervised learning algorithm with the cat folder and dog folder, the classifiers can try to label the images in the third folder. The algorithm will now label the unknown images as a cat or a dog image.

Unsupervised Learning uses a dataset with multiple features to learn from the features. Algorithms of this type aim to learn from the distribution of the dataset's vectors, $\mathbf{x}$, and try to extract interesting properties. The unsupervised learning algorithms do not get a label, $\mathbf{y}$, during training. This is the reason behind the name unsupervised. Let us say we have a folder containing both dog and cat images. This time, the folder is not labeled, and we do not know if it is a cat or a dog in the image. We convert each image to a simple vector and use them instead of the images. Now, using an unsupervised learning algorithm on the vectors shows two clear clusters of vectors. A cluster means a position in vector space where vectors are more clustered together. The two clusters are investigated where it is concluded that one corresponds to dogs and the other to cats. Now the images can be sorted based on which cluster the image vector is closest to. Now all images can be sorted into cat and dog folders, and everything was performed without knowing the content of the images.

Reinforcement Learning (RL) does not exclusively get any data directly. RL gets data by interacting with its environment and tries to learn what to do to become better [Goodfellow, 2016, p.105-106]. It creates it is own data by itself. Usually without knowing the mapping between actions and effects on the environment, the agent[10] attempts to discover which action returns the highest reward given the situation [Sutton, 2018, p.1-4]. For instance, let us say you want to make an agent that can play the video game Pong. The agents get as input the racket's[11] and ball's positions at all time. The actions the agent can decide to do is to move the racket up or down. The reward is the number of wins it gets. The agent will, over time, learn how to move the "board" and try to get higher rewards, but at the start still be bad. After many tries and errors, the RL agent will become be an great Pong player.

Furthermore, Semi-supervised learning is a field in ML that can be placed between Supervised and Unsupervised learning. Semi-supervised learning can be used to label data faster by combining elements from both fields. By giving a semi-supervised algorithm a small bath of labeled data and a larger non-labeled dataset, the system can learn similarities from the labeled data and use the knowledge to assign labels to the unlabeled data [Géron, 2019, p.13].

2.3.1 Passive Aggressive Classifier

Passive Aggressive Classifier is an online margin-based algorithm for binary or multi-class predictions. In machine learning, an online algorithm means an algorithm that does not need to see the whole data at once to learn from it. It rather learn step-by-step when the data come in sequential order. The methodology of the algorithm is based on generating hypotheses and then evaluating them to correct itself [Crammer et al., 2006]. It extracts knowledge from the "stream" of data and "discards" them afterward. The online aspect is beneficial in data of more real-time nature where data is a stream. The data can be used to train continuously or when the size of the dataset is large. The algorithm was published in the paper "Online Passive-Aggressive Algorithms"[Crammer et al., 2006] in 2006.

2.3.2 Support Vector Machine

Support vector machine algorithm is a supervised algorithm created in 1963 by Vladimir Vapnik and Alexey Chervonenkis. Vectors are represented in the vector

[10] An agent is something that can act with its environment based on what it sees in its surroundings [Russell and Norvig, 2016, p.4].

[11] The "movable" block is actually a tennis racket in Pong.

space where the algorithm divides them by using a margin line/surface, called an optimal hyperplane. It aims to find the maximum margin from support vectors. As illustrated in Figure 2.2, using the closest vectors to the margin called support

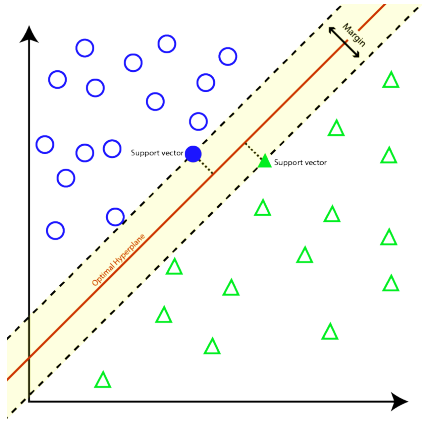

Figure 2.2: An illustration of a Support Vector Machine(SVM)

vectors, creates a section where the different types of points are distributed. The section and "line" make it possible to predict when a new point is introduced. During training, the hyperplane is rotated and changed to maximize the margin distance between the support vectors to adjust the hyper-plane better. Kernel trick are also used to map the inputs to a high-dimensional feature spaces, and use this for improving classification.

2.3.3 Decision Tree Learning

Decision Trees is an unsupervised ML and is an undirected graph with a root node, where child nodes are created based on the data's features [Mitchell, 1997, p.52-53]. The learning aspect is performed by selecting good features from the attributes to split based on a target value. A metric is used to determine "good" attributes to select as the next node in the tree—metrics such as Gini, information gain, and entropy [Mitchell, 1997, p.55-60]. The end structure has similarities with flowcharts, where the node features and the edges to the other levels are the threshold values [Mitchell, 1997, p.52-53].

However, decision trees tend to overfit, especially when the depth of the tree grows. A solution to this is the **Random forest**, where a collection of decision trees are used for a final result [Russell and Norvig, 2016, 698-707].

2.3.4 Logistic Regression

Logistic regression is one of the easiest machine learning algorithms. It is a supervised learning algorithm used for classification and regression. The ML model is based on the mathematical logistical regression model, and it is used i to find the probability of an example belonging to a class. The algorithm uses the sigmoid function on the output to deliver the discrete classes' likelihood [IBM, 2022].

2.3.5 Naïve Bayes Classifier

Naïve Bayes Classifier is based on the probability formula Bayes Theorem, and uses the theorem with a naïve assumption of conditional independence to pair together data **X** with class **Y**. *Class* means a label to a data, making it a part of a class. As shown in Equation 2.2, the formula solves the conditional probability P($A|B$) by multiplying the prior probability of event A happening with the conditional probability of event B given event A, divided by the prior probability of B.

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \tag{2.2}$$

Conditional probability is the degree of belief of an event happening given another event, meaning the system "updates" beliefs when presented with new cases.

- P(A): Probability of A
- P(B): Probability of B
- P($A|B$): Conditional probability of A given B
- P($B|A$): Conditional probability of B given A

With this as the foundation for the Naïve Bayes Classifier, the classifier estimates the most probable classification of an example given its features based on its "learned" belief degrees [Zhang and Su, 2008]. The classifiers can be extremely fast compared to more sophisticated methods[12].

2.3.6 Artificial Neural Network

Origin of Artificial Neuron

In 1949 a book named "The Organization of Behavior" by Donald Hebb introduced the Cell Assembly Theory, also known as Hebbian Learning [Hebb, 1949].

[12] `https://machinelearningmastery.com/naive-bayes-for-machine-learning/` [Accessed in 10.06.2022]

The Hebbian rule, based on the Hebbian Learning, summarises to "cells that fire together wire together" and means that cells that activate each other should increase their connection between them. This is solved by increasing the weight connecting both nodes. The rule created the fundaments for creating the first artificial neural nodes and imitating such behavior with mathematical equations. Frank Rosenblatt created the supervised machine learning algorithm Perceptron in 1958 [Rosenblatt, 1958] and constructed the Mark I Perceptron, implementing it into reality. The Perceptron represents one single neuron, as shown in Figure 2.3, capable of taking inputs and giving a binary classification as an output. The

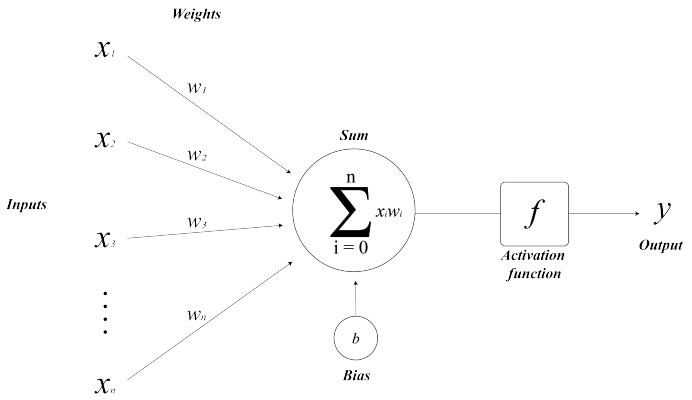

Figure 2.3: An illustration of the structure of a neural node

Perceptron changes the values of the bias and weights to improve the success rate of the classification. In the book [Minsky, 1988] the Perceptron was shown not to solve the XOR problem showing it's a limitation to non-linearly separable classification problems. This means problems of classification what cant be solved by one straight line. The book created the first "AI-winter" removing interest for the field but initiated the interest of finding a method of learning multiple layers of neural. The answer was an artificial neural network. A multi-layer Perceptron had the possibility to split a non-linear problem, but the process of adjusting the weights in the nodes was not implemented in artificial neural networks. Around the same time, the back-propagation algorithm was presented in the paper "Learning representations by back-propagating errors" [Rumelhart et al., 1986] introducing a mathematical method of training/adjusting multiple weights and biases in entire ANN.

Artificial Neural Networks(ANN)

ANNs is a interconnected collection of multiple artificial neurons, inspired by the structures of our brains. The complex process found in the brain is the inspiration to mimick how the brain works in computers to create "intelligent" behavior [Goodfellow et al., 2016, p.165]. The neural cell communicates by forwarding, or not forwarding, its signal with variation of strength. The strength in the artificial neural is referred to as weights. The anatomy of an ANN is usually divided into

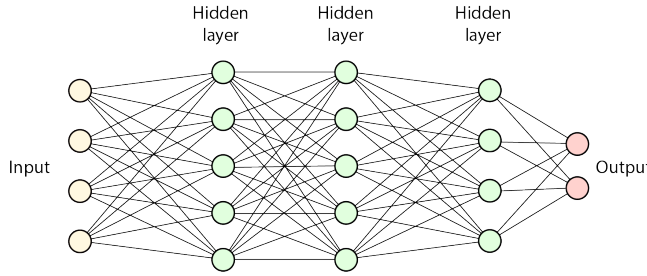

Figure 2.4: An Artificial Neural Network with three hidden layers

three parts: input, hidden layer, and output (as seen in Figure 2.4). The input is where the information is inserted into the system, while the output is the results from the neural network. The number of hidden layers can vary; more hidden layers create the possibility of solving more complex problems, but at the same time making the network more resource expensive to train. During the training of a neural network, it is presented with different examples and labels. This means the ANN is a supervised learner. The network passes the instance through all the neurons and compares the output with the correct label. The model then uses the algorithm backpropagation to readjust the weights. The idea is that the model will predict better when presented with unseen examples in the future. The re-adjustment of the weights and biases is the "learning" aspect in ANNs [Goodfellow et al., 2016, p.164-167].

Deep Learning (DL) is a sub-field in Machine Learning [Goodfellow et al., 2016, p.9] and is given the name *deep* duo to the number of hidden layers in the system giving the network greater depth [Goodfellow et al., 2016, p.164-165]. The available data produced on the internet and the increase of the computation processing made ground for Deep Learning to become an ever developing field in AI. The depth of the neural network makes it possible for the network to prioritise the important features by "itself" by rewarding features of higher weights. Sub-fields in DL where the network structure is different enhance, e.g., better results when working with images, learning from longer sequenced data

forms, or mimicking attention to better understand how to learn. Sub-field, such as Convolutional Neural Network (**CNN**), Recurrent Neural Network (**RNN**), **Transformers** (as BERT and GTP-3) are a part of DL.

2.3.7 Transformers

In 2017, the paper Vaswani et al. [2017] introduced the Transformer architecture. Before this paper was published, recurrent neural networks (RNNs), Long Short-Term Memory (LSTM), and gated recurrent neural networks were state-of-the-art in time series data. This is data where time or sequence is important, such as temperature, stock markets, and language. The "then" state-of-the-art models also accomplish good results in NLP tasks such as machine translation. They accomplished good results but still had some complications. The sequential nature of language made it challenging for the models to keep information on relationships between terms due to computational memory constraints. Additionally, training the models where time-consuming since they could not be trained parallel [Vaswani et al., 2017, p.1-2]. The Transformer architecture utilised the attention-mechanism and made it possible for parallelisation and to train in less time and with a lower cost [Géron, 2019, p.554]. Attentions allow the transformer to learn what it should focus on based on the input.

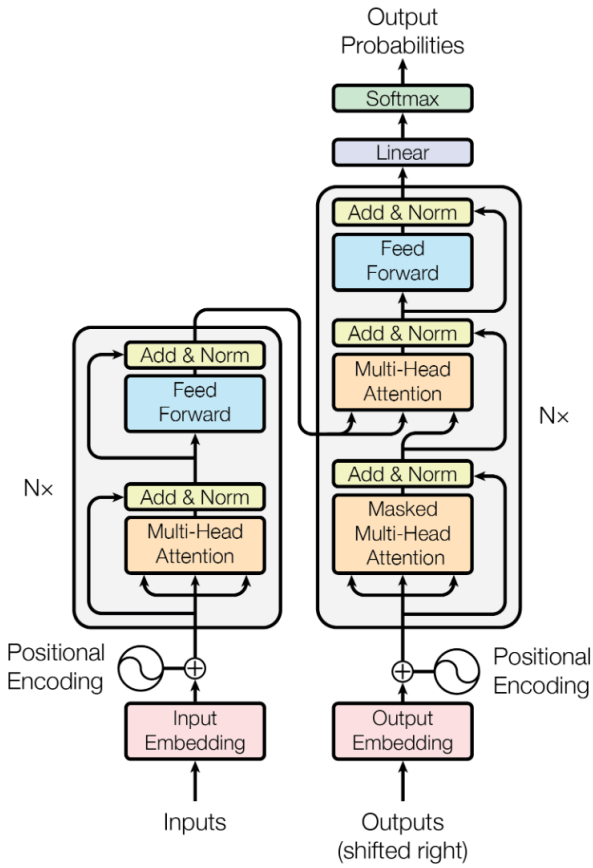

Figure 2.5: The Transformer architecture from Vaswani et al. [2017]. With permission from one of the authors, Llion Jones at Google.

The Transformer architecture has an Encoder-Decoder division where the left part is the encoder, and the right is the decoder (see Figure 2.5). The encoder takes the entire input sentence as a sequence of unique word IDs and converts them into a 512-dimensional vector representation. The decoder will, during training, receive a sequence of word ID from the targeted sentence word-by-word while being presented with the entire input sentence vectors from the encoder. The output from the decoder returns a probability of each next word [Géron, 2019, p.555-556]. This architecture includes well-established elements from ML as feed forward neural network and the softmax activation functions, but introduces further three new components; he *Multi-headed Attention, Masked Multi-headed Attention* and *the Positional Encoder*.

Multi-headed Attention and a *Masked Multi-headed Attention* components are added to the encoder and decoder to bind the different words together. The Multi-headed Attention component represents each word's relation with every other word in the sentence as an attention vector. The Masked Multi-headed attention does the same thing but differs by only giving it the attention before the current word in the sentence [Géron, 2019, p.556].

The Positional Encoding seen in the lower part of both the encoder and decoder (see Figure 2.5) represents the position of each word in a sentence with a vector. The purpose is to describe the sentence structure since the Multi-headed Attention components do "focus" on only the relations between the word. The positional encoded encodes the position of each letter by using a mathematical approach through the usage of the oscillating characteristics of cosine and sinus [Géron, 2019, p.555-558].

Together with other events, the Transformer gave an explosion of innovation in the field of NLP as the Generative Pre-training Transformer (GPT) and Bidirectional Encoder Representation from Transformers (BERT) [Géron, 2019, p.563-564].

BERT

The paper Devlin et al. [2018] introduced BERT to the world; The **B**idirectional **E**ncoder **R**epresentations from **T**ransformer. BERT is similar to the originally transformer architecture from Vaswani et al. [2017] [Devlin et al., 2018, p.3] but has changed and added components to tackle challenges the team detected. The model does not use traditional directional (left-to-right or right-to-left) language models. Instead, it uses a bi-directional language mode. This means it learns languages from left-to-right *and* right-to-left. During training, the original Transformer from Vaswani et al. [2017] shows the self-attention components only tokens

before the current token. In Devlin et al. [2018] it is argued to be harmful for the transformer model learning, and is a problem solved by introducing *Masked LM* and *Next Sentence Prediction (NSP)* [Devlin et al., 2018, p.3-5].

Masked LM aims to train a deep bi-directional representation by randomly (15% of the time) masking words in a sentence. To prevent the downsides of having a mismatch between the pre-training and fine-tuning of the model, additionally, "randomness" is included. When a word is "selected" from the sentence, 80% of the time, the word will be hidden with a *[MASK]* token. The other 10% of the time it will be replaced with a random token, and the remaining 10% of the time, it will be kept unchanged [Devlin et al., 2018, p.4].

Next Sentence Prediction sees and uses the importance of understanding the relationship between the sentences. Let's say that A and B are two sentences used during training. While training, 50% of the time, sentence B will be the following sentence after A. The remained 50% of the time, sentence B will be a randomly selected sentence from the corpus. The specific events are binary labelled as *IsNext* and *NotNext*. This is used during training [Devlin et al., 2018, p.4-5].

When the paper was published in 2018, BERT introduced a new method to approach NLP challenges and contributed to the field with state-of-the-art results to eleven different NLP tasks [Devlin et al., 2018, p.1].

2.3.8 Distance and Similarity

In mathematics and ML, similarity and distance are interchangeable terms. Latent space, called feature space or embedding space, will embed data to vectors and give them positions. If two positions are close, the points will share similarities. If they are very different from each other, the distance will be greater. This section will present methods to calculate the similarity and distance of different data types.

Euclidean distance

Euclidean distance is the length between two points in a euclidean space. The distance between the point is calculated using the Pythagorean equation. As seen in the Figure 2.6, the straight line between the red and orange is the distance euclidean distance.

$$distance\,(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

The shorter the distance between the points is, the more similar they are. If the distance is 0, meaning the points are in the same position, then they are identical. And if the distance is more significant, then the differences are also greater.

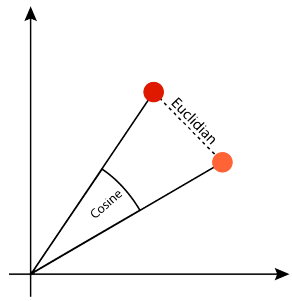

Figure 2.6: Euclidean-, and Cosine similarity

Cosine Similarity

Cosine similarity is a method of comparing two different vectors. The range in this metric goes from 0.0 to 1.0. From a mathematical aspect, the similarity is measured by finding the cosine degrees between the two multidimensional points in space.

$$\cos(\theta) = \frac{\tilde{\mathbf{V}}_\mathbf{a} \cdot \tilde{\mathbf{V}}_\mathbf{b}}{\|\tilde{\mathbf{V}}_\mathbf{a}\| \|\tilde{\mathbf{V}}_\mathbf{b}\|}$$

As seen in Figure 2.6, the angle between the red and orange plot is the cosine similarity. One in similarity means identical, while zero means very different.

Levenshtein distance

Levenshtein [1965] intrudes the world for the idea of Levenshtein distance. Levenshtein distance is referred to as the editing distance and is a metric to measure change between two strings of text. The method contains character-level operations as insertions, deletions, and substitutions of characters, where the aim is to find the minimal operations needed to get from one string to another. The sum of all processes necessary is the Levenshtein distance. A small Levenshtein distance represents that the two strings are very similar, while a higher number means the need for multiple operations and, therefore, very different from each other. E.g., the word "book" and "rookie" have a distance of 3, there "b" is substituted with "r" ("book" to "**r**ook"), "i" is inserted in the front ("rook" to "rook**i**"), and "e" is inserted in the front ("rooki" to "rookie").

2.4 Evaluation Metrics

This section stems mainly from the preliminary study and contains minor changes as mentioned in the introduction of the chapter. The original section can be found in [Ramirez, 2021, p.19-20]. Section 2.4.4 is an inclusion to this chapter.

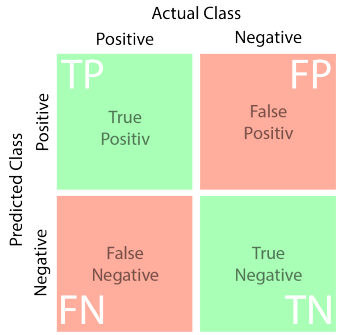

Figure 2.7: The Confusion Matrix

As seen in Figure 2.7, the abbreviations TP, FP, FN, and TN represent True Positive, False Positive, False Negative, and True Negative.

2.4.1 Precision

Precision can be understood as "how many of the predicted elements are correct?". It measures the model's relevancy in the results. High precision is desirable but cannot fully represent the system's performance. That is where recall can help [Géron, 2019, p.91].

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

2.4.2 Recall

Recall, called sensitivity or *true positive rate*, tells "how many of all the correct answers were predicted correctly". The recall represents how many predictions that were corrected based on all true positives and the false negative (the one falsely classified) [Géron, 2019, p.91].

$$Recall = \frac{TP}{TP + NP}$$

2.4.3 F1-score

Precision and recall together tell the performance of the systems. E.g., a high precision can result from prediction only one element correct, $\frac{1}{1}$, but this would be represented in the recall with a low score. To reflect this trade-off, a harmonic means utilising both precision and recall, creating a formula where both need to be high to acquire a high F1-score.

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{TP}{TP + \frac{FN+FP}{2}}$$

In the F1 score, both precision and recall are favored similarly [Géron, 2019, p.92-93].

2.4.4 K-fold Cross Validation

K-fold cross-validation is a method of estimating the performance of a system. The idea behind the k-fold is to divide the data into k numbers of approximately equally sized folds. The system is then tested iterative k times by holding off on fold, k, and trained with the remaining folds, k-1. Each iteration uses the "selected" k-fold to test on and the remaining fold to train with. In this manner, the results from the testing are averaged by the number of iterations/folds to better measure the performance of the system [Yadav and Shukla, 2016, p.80]. For instance, if a classifier is cross-validated 10-fold, the data is divided into ten equally sized segments. For the first iteration, the system selects the first segment to test on and the rest to train. The second iteration continues by selecting the second segment for testing and the other segments for training. The iteration is executed ten times before the results are averaged.

The benefit of using k-fold cross-validation is that it provides a better overview of the model. The results give an overview of its performance when presented with unknown data. The disadvantages are that the model is run multiple times. This can be resource expensive and time-consuming.

2.5 Frameworks and Libraries

This section presents the important frameworks used for gathering data and analysis of the data. As mentioned in the introduction in Chapter 2, this section is new and not from the preliminary study.

2.5.1 Garc and Tweepy

Garc is a library that simplify fetching of information from the platform Gab. The code is public on GitHub[13] and through Python's package index manager, PIP. Garc has multiple API wrapping functions, where the extraction of entire user-profiles and all their posts was utilised during the creation of the radical dataset (see Section 4.2.2).

Tweepy[14] is similar to Garc but instead retrieves information from Twitter. It is an API-wrapping library that simplifies the task of retrieving information such as user profile information or user post. Configuration is also straightforward by adding the different tokens and secret keys acquired after creating a Twitter Developer account. The library had an essential role in the creation of the non-radical dataset (see Section 4.2.3).

2.5.2 Beautiful Soup

Beautiful Soup[15] is a library that helps manipulate HTML and XML files and retrieve information from the files. It has powerful features making parsing, searching and modifying the parsed tree simple. It enables Python to work with files from websites by converting the file to a Python dictionary/hash-map.

2.5.3 KeyBert

KeyBERT [Grootendorst, 2020] is a term extractor for retrieving the best keyword to represent a document. It utilises the embeddings of the Transformer BERT (see Section 2.3.7) to embed both documents and the term and compares the different word(s) to the entire document. First, it creates a set of "candidates" by selecting the most frequent words. The document is vectorised and compared to each vectorised candidate with cosine similarity to find the most similar ones. The assumption is that the candidates with the highest similarity similarity will be the best keywords to represent the original document. The system returns all the candidates in descending order based on the cosine similarity. KeyBERT also has two measurements for creating a diversion in the different terms; the Maximal Marginal Relevance and the Max Similarity measurement [Giarelis et al., 2021, p.638-639].

The creator of KeyBert aimed to create an easy-to-use implementation and was inspired by other projects for the BERT approach to keyword extraction. As seen

[13] `https://github.com/ChrisStevens/garc`
[14] `https://github.com/tweepy/tweepy`
[15] `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`

in the comparative study Giarelis et al. [2021], the keyword extraction approaches TF-IDF, Rake, YAKE, KeyBERT, TextRank and SingleRank were compared with six different datasets. The study concluded that KeyBERT performed best with longer text and was more suitable for data of that nature [Giarelis et al., 2021, p.643-644].

2.5.4 Scikit-learn

Scikit-learn[16] is a library with unsupervised and unsupervised learning algorithms. It is for the coding language Python and contains popular algorithms such as support-vector machine, k-means, random forest, and more. The machine learning classifiers used in the experiments in Chapter 4 are from this library.

[16] `https://scikit-learn.org/`

Chapter 3

Related Work

This chapter presents the process of acquiring knowledge of the domain. The methodology was based on the paper Kofod-Petersen [2015] and is explained in detail in the following sections. The different sections represent the different phases of executing the structured literature review (SLR) in Kofod-Petersen [2015]. Section 3.1.1 presents the planning phase of SLR. Later comes Section 3.1.2 where the conduction phase is explained. The same section demonstrates the different queries used for finding literature about radicalisation. The last parts of the chapter in Section 3.1.3 present the exciting and relevant findings sorted into groups/subsections. All the findings in Section 3.1.3 stem from the literature from the SLR.

The SLR was executed from August to December of 2021 for the preliminary study as presented in Section 1.2. Section 3.1 and 3.1.3 are retrieved from the preliminary study. Minor corrections have been made, such as structure and grammatical errors.

3.1 Structured Literature Review (SLR)

The preliminary study aimed to acquire an overview of the area of radicalisation. The research included particular traits that could be found among individuals before radicalisation. Additionally, the study discovered various methods on finding the radicalisation point: the point where a person is radicalised. The method selected for the task was, as mentioned, SLR from Kofod-Petersen [2015]. The execution followed the paper by dividing the steps into three parts: planning, conducting, and reporting.

3.1.1 Planning

The primary purpose of conducting a SLR is to learn about the domain of interest. Finding literature that covers the domain more systematically is beneficial as it reduces biases and proposes a reproducible method for the research. Planning consists of finding the need for SLR, creating the research question, and then "reviewing" protocols for later processes. Conduction primarily consists of actively narrowing down the pool of research papers. The search was performed in Scopus, a search engine service provided by the company Elsevier to find academic papers [1]. The service contains a range of searching tools for finding relevant papers, including citation detection and search with logical operators. Research from 2009 that compared Web of Science, Scopus, and Google Scholar found that Scopus had high quality search results with comprehensive coverage of international papers. Also, according to the developing team, Scopus updates its articles daily [Kulkarni et al., 2009]. Additionally, the Scopus search engine allows the researcher to decide where the search is conducted, e.g., "Article title", "Abstract" and "Keywords", which assures higher quality results compared to other search engines.

3.1.2 Conducting

Step 1: Identification of research

To find relevant literature using the SLR, it is crucial to select terms that effectively cover the goal and research questions. The initial step of searching is to discover how to break down the field of search into groups of terms. The following four groups were created:

- **Group 1**: Related to the act of changing or affecting a person's behaviour.
- **Group 2**: Related to terms of how to discover radicalisation.
- **Group 3**: Related to social media platforms.
- **Group 4**: Related to the computational aspect of the research (Computer Science and Artificial Intelligence).

By uniting the groups with "AND"(conjuncture) and uniting the terms inside the groups with "OR"(disjunction), a query can be made. The idea is that the relevant research papers will contain at least one term from each group. Created by the words in Table 3.1, the initial query became:

[1] `https://www.elsevier.com/solutions/scopus`

	Group 1	Group 2	Group 3	Group 4
Term 1	Radicalization	Traits	Twitter	Machine Learning
Term 2	Extremism	Indicators	Facebook	Artificial Intelligence
Term 3	Brain Washing	Detection	Reddit	Natural Language Processing
Term 4	Radicalisation	Prevention	Gab	
Term 5		Vulnerable		

Table 3.1: Search terms sorted into groups

1th Query: *TITLE-ABS-KEY*[2] *((Radicalization OR Extremism OR "Brain Washing" OR "Radicalisation") AND (Traits OR Indicators OR Detection OR Prevention OR Vulnerable) AND (Twitter OR Facebook OR Reddit OR Gab) AND ("Machine Learning" OR "Artificial Intelligence" OR "Natural Language Processing" OR "AI" OR "ML" OR "NLP"))*

Modifications to the fourth group were made by adding different abbreviations of the fields to the query. Artificial intelligence, machine learning, and natural language processing are commonly written as *AI*, *ML*, and *NLP*. The search yielded 17 results with no specific times set. At first glance, the research articles show promising titles and high in relevance. They range from being published between 2015 and 2022. All the articles were collected and selected to go forward in this process.

2th Query: *TITLE-ABS-KEY((Radicalization OR Extremism OR "Brain Washing" OR "Radicalisation") AND (Traits OR Indicators OR Detection OR Prevention OR Vulnerable) AND (Twitter OR Facebook OR Reddit OR Gab)*

The second query is a modified version of the first query. The modification includes the removal of the fourth group and was decided through conversation with Björn Gambäck. His recommendation was not to exlude other fields of science. Research from other fields like social studies and psychology can also be relevant for this thesis.

3nd Query: *TITLE-ABS-KEY(("Jihad" OR "Jihadism" OR "Terrorism" OR "Islamic radicalisation" OR "Far-right" OR "Far-Left" OR "Fascism") AND (Traits OR Indicators OR Detection OR Prevention OR Vulnerable) AND (Twitter OR Facebook OR Reddit OR Gab) AND ("Machine Learning" OR "Artificial Intelligence" OR "Natural Language Processing" OR "AI" OR "ML" OR "NLP"))*

[2]Scopus' "Article Title, Astract and Keywords" search command: `http://schema.elsevier.com/dtds/document/bkapi/search/SCOPUSSearchTips.htm` [Accessed in 10.06.2022]

The third and last query is a modified version of the first query, where the first group of terms were changed to more specific radicalisation types. After searching through literature surrounding the topic, political and religious radicalisations dominanted the field. Terms like "Jihad", "Jihadism", "Islamic radicalisation", "Terrorism" and "Far-right", "Far-left", "Fascism" were added to the query. This search gave the result of 19 new papers adding up to the total sum of *83 papers*.

Step 2: Selection of primary Studies

After performing the search, the total size of the return articles was 83 papers. The aim was to reduce the size to a manageable size of 15-25 papers. Protocols to assure quality in the final literature list are explained later in the process, but to initiate the reduction process, the following list of protocols were used.

- Duplicates.
- Same study published in different publisher.
- Studies are limited to a certain time range.

Removal of duplicated articles eliminated 22 papers, making the total amount of 61 papers. Articles published multiple times by different publishers removed an additional two papers, making the total sum of 59. The time range of the documents ranged mainly between 2015 and 2022, in which only one was from 2013. The last measure brought no new results, making the total sum of 59 papers. Additionally, eight papers were not accessible without permission from the author. After no responses from the authors, they were removed from the set, reducing the total number of papers to 51.

Step 3: Study quality assessment

Primarily Inclusion Criteria (IC)

The method of identifying whether criteria (see Figure 3.2) IC1, IC2, and IC3 is fulfilled is by reviewing the papers' titles and abstracts. IC1 assures that the paper is in the domain of radicalisation with other similar domains, including hate speech, extremism, violence and political radicalisation. IC2 secures that the paper uses AI, while IC3 assures that the paper is an experiment or a study containing an overview of other papers. After the first step, the total number of articles was 33.

Secondary Inclusion Criteria (IC)

The Secondary Inclusion Criteria included criteria IC4 and IC5. IC4 assures that the paper is within the specific field of detection of traits and people who

Criteria Identification	Criteria
IC1	The study's focus is extremism or radicalisation on social media.
IC2	The study is in the field of AI.
IC3	The study presents empirical results.
IC4	The study focuses on the process of radicalisation or detecting people vulnerable to radicalisation.
IC5	The study focuses on the use of NLP and text processing.
QC1	Clearance on the aim of the study.
QC2	The study uses other similar research into consideration of its research.

Table 3.2: Criteria: Inclusion Criteria (IC) and Quality Criteria (QC)

are vulnerable to radicalisation. If the research papers did not include these criteria, it was essential to ensure that they were at least within the scope of radicalisation. In order to prevent the removal of a significant number of papers, it was essential to be somewhat flexible with the criteria. For instance, there has been conducted way more research on the detection of radical people rather than early indicators of radicalisation. Thus, to avoid the exclusion of all results showing papers about radicalisation, some flexibility was deemed necessary. For instance, it was deemed likely that the method and theory parts of some papers would include the detection of traits or signs of vulnerabilities to radicalisation. The second criteria, IC5, secures that the paper uses methods from NLP. In some cases, if the proposed process included an innovative method not specifically in the field of NLP, the criteria could possible be ignored. Combining the methods with other approaches could potentially be something to look into. The process of controlling whether the paper passes the criteria is performed by skim-reading through the papers. The total size of the papers was reduced from 33 to 25.

Step 4: Quality Criteria

After searching and controlling the papers with both the Primary-, and Secondary Inclusion Criteria, the remaining articles were 25. The next step was to award points to the reports based on the Quality Assessment Criteria (QAC). The QAC is presented in Appendix A, and is a modified version of the QAC presented in Kofod-Petersen [2015]. The point-range goes from zero to one, where zero means the paper does not fulfill the criteria, while one means it does. Additionally, a half-point is possible if the article partly fulfills the requirements. This award system was the most consistent and discrete. The results from the assessment

are presented in Table 3.3.

Step 5: Data Extraction

The papers underwent a data extraction based on the listed fields in Appendix B.

Step 6: Data Synthesis

Step 5 produced the following results and are presented in Table 1, Table 2, and Table 3. The tables are all in Appendix D. The columns in the table correspond with the listed points in Appendix B.

3.1.3 Results of the Structured Literature Review

The results after performing the literature review are presented in three different tables in Appendix D; Table 1, Table 2 and Table 3. The rest of the chapter presents the findings from the SLR.

Findings

Index	Referance	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Sum
1	[Agarwal and Sureka, 2015]	1	0	0	1	0	1/2	1	1	1/2	1	1	7
2	[Alghofaili and Almishari, 2018]	1	1	1/2	1	0	0	1/2	1	0	0	1	6
3	[Aljarah et al., 2021]	1	1	1/2	1	0	0	1	1	1/2	1	1	8
4	[Alvari et al., 2019]	1	1	0	0	0	0	1/2	1/2	1/2	0	0	3.5
5	[Benigni et al., 2017]	1	1	1	0	0	0	0	1	1/2	1/2	0	5
6	[Chatzakou et al., 2020]	1	1	1	1/2	0	1/2	1	1	1	1	1/2	8.5
7	[Chelvachandran and Jahankhani, 2019]	1	1	1	1	1/2	1	1/2	0	1	1	1	9
8	[Deb et al., 2020]	1/2	1/2	1/2	1/2	0	0	0	0	1/2	0	1/2	3
9	[Derbas et al., 2020]	1	1	1	1/2	1	0	1/2	1	1/2	0	1	7.5
10	[Fernandez and Alani, 2018]	1	1	1	1	1	1	1/2	1/2	1	1	1	10
11	[Fernandez et al., 2018]	1	1	1	1	1	1/2	1	1	1	1	1	10.5
12	[Ferrara et al., 2016]	1	1	1	1	1/2	1/2	1	1	1/2	1	0	8.5
13	[Grover and Mark, 2019]	1	1	1	0	0	1/2	1	1	1	1/2	1/2	7.5
14	[Hartung et al., 2017]	1	1	1	1	0	1/2	0	1/2	1/2	0	1/2	6
15	[Kostakos et al., 2018]	1	1/2	1/2	1	1/2	1/2	1	1	1	1	1	9
16	[Lara-Cabrera et al., 2019]	1	1	1	1	1	1	1	1/2	1	1/2	1/2	9.5
17	[López-Sáncez et al., 2018]	1	1/2	1/2	1	1/2	1/2	1	1/2	1	1/2	0	7
18	[Miranda et al., 2020]	1/2	1	1/2	1	0	1/2	1/2	1	1	1/2	1	7.5
19	[Necaise et al., 2021]	1	1	1/2	1	0	1/2	1	0	1	1	0	7
20	[Nouh et al., 2019]	1	1	1	1/2	0	1/2	1/2	1	1	1	1	8.5
21	[Oussalah et al., 2018]	1	1	1	1	0	1/2	0	1	0	0	1	6.5
22	[Rekik et al., 2020]	1	1/2	1	1	0	1/2	1	0	1	1/2	1/2	7
23	[Saif et al., 2017]	1	1	1	1	1/2	1	1	1	1	1	1	10.5
24	[Saif et al., 2016]	1	1/2	1/2	1	1	1	1/2	1	1	1/2	1	9
25	[Ul Rehman et al., 2020]	1	1	1	1	1/2	1	1	1	1	1	1	10.5

Table 3.3: Scores from quality assessment

Based on the QAC, here are the points assigned to each paper after performing the quality assessment. The criteria from QAC are represented in each of their own columns in Table 3.3. The criteria can be found in Appendix A.

3.2 Feature Selections

Deciding the features to train a machine learning model is half experimenting and half gut-feeling. Which features contain sufficient discriminatory power so that a machine learning model can improve its prediction? Some papers had an analytic approach to features, where a "kitchen-sink" method was used to find the best features based on the results. Others stumbled over the findings of features by "accident". The following section will present interesting findings about features used for detecting or predicting radicalisation.

3.2.1 Interesting Findings

Features used in a detection/prediction model vary. In Ferrara et al. [2016], a proposed machine learning framework was trained to predict and detect radicalisation with 52 different features. All the features represented the entire user

profile and corresponded to three classes: user information, timing features, and network statistics. The first class contained user-related features of the user's profile and meta-textual information of the tweets. The meta-data included number of followers and friends, ratio between retweet and tweets, and so on. The timing features included temporal data that focused on features like the average number of tweets per day. The network class was about the interaction between the followers. Three tasks regarding the extremist support detection and interaction were experimented upon, and the overall result was analysed together with the most important features. Overall, the top ten features ranking from best to worst went from the ratio of retweets/tweets, the average number of hashtags, number of tweets, average number of retweets, tweets per day, the average number of mentions, number of followers, number of friends, average mentions, variation in tweets per day, and the ratio of mentions/tweets.

In Lara-Cabrera et al. [2019], psychological characteristics can be observed in textual utterances, making it possible to extract psychological traits about individuals. This field is called psycholinguistics. Two features are mentioned as detectable indicators of radicalisation: introversion and frustration. Introversion is shown through short sentences, while frustration can be seen through the percentage of profanities used in the text. The initial hypothesis was based on the presumption that radical users used more swear words and wrote shorter sentences. However, the results supported the theory on frustration, but not introversion. The final results showed that radical users wrote longer than common users.

In Agarwal and Sureka [2015], they found that internet slang, emoticons, and punctuation are essential features that can indicate hateful content. The presence of religion-related terms, war-related terms, offensive words, and negative emoticons were solid indicators of hateful content related to Jihad.

Fernandez et al. [2018] argue that only using term-based detection for radical content may give false positives. This may happen because other Twitter users report on radical events, express harmless religious utterings, and even make utterances that condemns extremism. Combining meta-data features with textual information can result in higher accuracy for prediction or detection. These results are coherent with the results of Fernandez et al. [2018], which showed an F1-score of 0.901 for detection and F1-score of 0.7-0.8 for prediction. The approach used the theory of "Roots of Radicalization" from social sciences.

In Deb et al. [2020], they used synthesised WhatsApp data to perform their research. They concluded that communication from riots/terrorist groups hap-

pens multimodal and not only through text. Especially in WhatsApp, images and voice recordings were used to provoke mass enragement. The authors suggested that future research should investigate the use of computer vision on images to improve detection.

In the paper Nouh et al. [2019], a language model (LM) was created by using the English magazine Dabiq (Pro-ISIS magazine) and used to create a psychological and linguistic profile. The profile was used to detect tweets supporting ISIS. The approach was innovative by including a psychological profiling element for detection. The psychological profile was created by using the LIWC dictionary. This dictionary contains words with corresponding emotions. The terms were used to train a classifier to detect the psychological profiles of users. The radical corpus was created by using TF-IDF and word2vec. Word2vec was better at finding important terms. The best psychological and linguistic features were radical psycho-profile distance, "Us" and "Them" frequency, number of mentions a user make of other users, user rank (based on graph-influence), sad emotions, etc.

3.2.2 Extraction Semantic Information and Other Metadata

In the paper Saif et al. [2017], textual data is combined to create a semantic graph-based approach. The purpose it to discover the semantic features with help from DBpedia. Terms are analysed by using DBpedia, and the features discovered were used in the detection task.

- Extracting name entities and their semantic context in the tweets.
- Build the semantic graph for each user representing concept and semantic relationship between the concepts.
- Apply frequent sub-graph mining on the semantic graphs to capture patterns to detect discriminatory patterns.
- Use the patterns as a feature for an ML classifier.

The usage of semantic features in the classification models resulted in an average of 7.8% average improvement to the F1-score. In another publication by Saif et al. [2016], they investigated the potential of including semantic information in classification. The semantic features are extracted with AlchemyAPI from DBpedia, YAGO, OpenCyc, Freebase, etc. Compared to the other non-semantic features, the extracted features improved the system. Together with network-data, the new features improved the system by 2% in the F1-score.

The study in Kostakos et al. [2018] focused on "The Manchester shooting" and "Las Vegas shooting". The study analyses the tweets surrounding the events with sentiment analysis, topic analysis and fake news detection. By using terrorism-related terms, such as ISIS and terrorism, they were able to recover tweets that were published around the time of the terrorist attacks. Semantic analysis was conducted using two methods, one with NLTK and SeniWordNet, and the other with SentiStrenght. Both methods were tested on the Sanders Twitter Sentimental Corpus. The results from the system improved when "neutral" tweets were changed to either be labeled as containing negative or positive sentiments. Topic analysis was performed by using Gensin' Latent Dirichlet allocation. The fake news detection was performed with a pre-trained Passive-aggressive (PA) classifier. The classifier was only used on the titles of the URLs during this study. Sentiment analysis showed an increase in negative tweets and echo chambers. The topic detection used Jensens-Shannon's divergence metrics to calculate topics discussed in echo chambers during the terror events.

3.3 Methodology

3.3.1 Detection

Classical and ML Approaches

In Agarwal and Sureka [2015], after preprocessing a Twitter dataset and labeling the tweets with help of a semi-supervised learning algorithm (KNN), SVM was used to classify the presence of hate-promotion. The KNN estimates the label by using the 100 nearest neighbours.

The approach in Aljarah et al. [2021] used preprocessing by removing non-Arabic words, weblinks, hashtags, symbols, numbers, diacritics, stop words, and negation words before utilising TF-IDF, BoW, or term-frequency. 15 different combinations of the emotion- and profile features were tested on four classifiers. Training with SVM (LibSVM) and Random-Forest (RF) were performed with GridSearch for hyper-parameter tuning. The last two classifiers were Naïve Bayes and Decision Trees. The model with RF achieved the best accuracy and presented a method that can be utilised regardless of language.

The research in Miranda et al. [2020] focuses on the detection of radicalisation related to ISIS in Indonesia. The usage of preprocessing to clean and tokenise the tweets continues with TF-IDF. The terms got their unique weights, making it possible to train an SVM. The results show an accuracy of 0.83 and show promising results in detecting radicals in the Indonesian language. This also indictes

tha the language of the tweets are irrelevant as long as the entire corpus is the same language.

In Alghofaili and Almishari [2018], they found that there was a need for automation due to increasing traffic from ISIS sympathisers on Twitter. They collected tweets from 600 accounts with an average of 3,200 tweets each, where 100 were terror incentives profiles. The non-inciting accounts were labeled as "well-known" in sports, politics, economy, health, and art, and the rest as religious or unknown. The algorithms Random Forest, Naive Bayes, J48, SVM, Naive Bayes Multinomial were compared to each other with accuracy and recall. Random forest displayed the best results. Potential changes for the future would be using information gain and n-grams.

Keyword Detection

Creating a dictionary of the terms related to the target scope is challenging and important when using term-related research. In some approaches, term frequency or term occurrence in text are used to detect radical users/content. As in López-Sáncez et al. [2018], terms were created by a domain expert. The terms were used to search for extreme users by their tweets by using logical operations. This refers to logical operators as "AND", "OR", "NOT", and such. The method helped the authors find extreme users. A less human-dependent approach is seen in Ul Rehman et al. [2020], where TF-IDF weighting was used to find highly weighted terms from one of the datasets. The research had a "radical" corpus containing 17350 tweets from 112 pro-ISIS users and used TF-IDF with n-grams. This is because previous work had shown improvement in detecting hateful speech with n-grams. The high weighted n-grams were then analysed to see differences between radical and non-radical users.

Username-based Detection

An exciting and additional approach for detecting radical users (or rather to rediscover users) is the proposed method by Alvari et al. [2019]. The results analysed username patterns found in radicalised and regular users. Using Levenstein distance on both groups shows that radical users tend to use more similar patterns. These findings can help to re-detect new accounts from previously banned users of the social media service or detect if a radical person is managing multiple accounts. Features like unique characters and the length of the username were concluded to be the best features for detecting radical users.

Mathematical Detection

In Oussalah et al. [2018], a metric is proposed for calculating radicalisation. The metric calculated the radicalisation score and then sent it to a KNN-SVM machine learning model. The ML approach includes n-grams, personality traits, emotions, linguistics, and network-related features. The dataset used stemmed from Twitter and Tumblr. The score is inspired by the typical approach to measuring the semantics of tweets. The radical score accounts for:

- Average sentiment score percentile (AS), given by the average sentiment score for all posts.
- The volume of the negative post (VN) is the percentage of negative posts in relation to all posts by the user.
- The severity of negative posts (SN) is calculated as the percentage of the very negative posts in relation to all negative posts.
- Duration of negative posts (DN), calculated based on the first day a negative post was posted and the last day.

The paper does not have a clear conclusion, but the table of results show good results overall.

3.3.2 Prediction

The article Fernandez et al. [2018] presents a computation approach for detecting and predicting the radicalisation influence on users based on the social science model "Root of radicalisation". The micro, meso, and macro effect were measured with the proposed computational method to measure radicalisation (see Section 3.5). The theoretical method was tested on "How ISIS uses Twitter" and showed promising results. Additionally, the prediction aspect of the approach was made by using the Collaborative filtering (CF) method. CF makes it possible to fill in empty values based on similar users and calculate the users' empty value.

In Lara-Cabrera et al. [2019] features were analyses to find the best features for detecting radicalisation. The features investigated were based on studies of traits commonly found in radical people. Five features belong to personality traits, and the rest are based on indicators of attitudes and beliefs. During the development of the research, two traits were selected; introversion and frustration. The feature introversion was based on the average length of the sentences and rooted in the presumption that introverted people usually express themselves in shorter sentences. Frustration was measured with the ratio of profanities and "normal"

words. Frustration is manifested in the more frequent use of profanities. Results were defined as "generally well", supporting the metric for frustration with statistical proof, but introversion was found to be the contrary. In fact, radical users tend to write longer sentences.

The research in Necaise et al. [2021] focuses on users' communication changes and how they can measure radicalisation. Potential features can be extracted from a user's textual tweet timeline and find patterns to predict the presence of the radicalisation process. In conclusion, highly active users (multiple radical subreddits) are more novel in commenting over time, meaning more changes than other medium-low ranking radical users. These findings can indicate potential traits for both detection and prediction of radicalisation.

3.4 Dataset

As mentioned in Fernandez et al. [2018], the quality of the dataset can be challenging since a golden standard for a radical dataset does not exist. Usually, datasets are not quality assessed by an expert and only created by unprofessionals. Quality control can assure that the dataset do not contain false positives or other misleading information. Here are the most reoccurring standard datasets used in the field, and different approaches used to create datasets.

3.4.1 Reddit Origin

The dataset used in the Grover and Mark [2019] was acquired from the subreddit "r/alt-right", a community focused on the alt-right ideology. The dataset was collected through a time period of 6-months until banned from Reddit. They utilised the *pushshift.io* API and extracted all the comments' data from the site. They extended the number of subreddits targeting politically oriented subreddits to later perform a textual term analysis. The other politically oriented subreddits were "r/conservative", libertarian", "r/democrats", "r/republican", "r/progressive", "r/socialist", "r/anarchist, "r/anarchocapitalist". The approach utilises subreddit communities and can construct an adequate dataset by focusing on the "democratic" powers of upvoting. They can indicate a post's relevance for the subreddit's topic.

3.4.2 Twitter Origin

The papers from the SLR concluded that the dominating origin of data was from Twitter. As shown in Table 1, Table 2, and Table 3 in Appendix 3.1.3,

papers with data originating from Twitter can be found in 22 of the 25 articles. Twitter-originating data can be found in both datasets available or as the preferred data source when collecting data. Here, both "famous" datasets and different approaches to collect data will be presented.

Dataset: "How ISIS uses Twitter"

Used in Fernandez et al. [2018], Fernandez and Alani [2018], Lara-Cabrera et al. [2019], Nouh et al. [2019], and Ul Rehman et al. [2020], this is the dataset most mentioned throughout the structured literature review. The dataset has been used for both detection and prediction and is a popular dataset on the site Kaggle. The public dataset "How ISIS Uses Twitter"[3] contains over 17,000 tweets from over 100 pro-ISIS users dating back to Paris Attacks in 2015. As described on the Kaggle, the attributes in the data are:

- Name
- Username
- Description
- Location
- Number of followers at the time the tweet was downloaded
- Number of statuses by the user when the tweet was downloaded
- Date and timestamp of the tweet
- The tweet itself

The dataset is created by a non-profit governance agency called Fifth Tribe. On the Kaggle page, there are multiple suggestions for using the dataset. The suggestions are sentiment analysis, network clustering, keyword analysis, and more [Fifth Tribe, 2019].

SAFFRON European project

The dataset created in the paper Derbas et al. [2020] originate from Twitter. It was created by using a crawling tool called Safapp. The tool is an outcome of the SAFFRON European project that focuses on detecting recruitment by terrorist groups online. The dataset was constructed by using the SafeApp finding "kill-events". Kill-events are sentences containing an event involving killings. The sentences contain arguments of killer, place, time, cause, instrument, and means.

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

An example given in the report of a kill-event is: "In April 2014 (time), there were reports that al-Asiri (victim) had been killed (kill)". The elements from the kill-event were used in the system to improve the results but still had room for improvement.

3.4.3 Snowballing Method

A snowball method for creating a dataset can be observed in Hartung et al. [2017], where the aim is to create a right-wing extremist dataset. As said in the paper, 37 seed profiles of political actors from the state Thüringen in Germany were used to create a radical and non-radical dataset. The 37 consisted of 20 profiles labeled as right-wing and 17 as non-extremist. The seed users' timeline and their followers were extracted with the Twitter API. A classifier was trained on the seed users and tested on a set of 100 random followers. The followers are extracted from the followers of the seed profiles, meaning people following the seed users. The testing used 100 randomly selected followers to the seed profiles. The idea was to see if the classifiers could detect the followers as radical or not radical. The 100 followers were labeled by a domain expert beforehand such that the evaluation of the test could be performed. A similar approach can be seen in [Benigni et al., 2017] to create a dataset, where the research plan is to use graph-based algorithms and clustering to detect online extremist communities. The dataset was created with initial five seed users related to ISIS. The following connections by the five users were analysed, and created a set of 1345 unique followers. Then, they extracted all the user profile information. The results ended up with 119'156 users and 862 million tweets.

3.4.4 Hashtag-based Method

Hashtags in tweets make it possible for a user to "label" their tweets. A user can either use a previously created hashtag or create a new one. The hashtags are characterised with an "#" in the front of the term. Another result is the opportunity to search and locate tweets where a specific hashtag was used. In Alvari et al. [2019], a dataset with approximately 1.6M tweets was created by using hashtags. The hashtags were predefined and related to radical content and, in this case, these were Jihadism and ISIS. The terms, as presented examples in the paper, were #AbuBakralBaghdadi, #ISIL, #ISIS, #Daesh, and #IslamicState. The researchers labeled 150 suspended ISIS-related tweets. The tweets were used as positive instances of radical content. Negative instances were created by randomly selecting tweets. One of the researcher's questions focused on evaluating the possibilities to correctly label violent extremists based on their Twitter username with supervised and semi-supervised learning. The semi-supervised method was given 150 labeled users with different features. The supervised was trained

on both negative and positive samples. SVM got the highest precision(0.96), and LabelSpreding and Char-LSTM achieved the highest F1-score. The overall outcomes showed promising labeling results to predict unseen labels.

In some research, the process moves from already existing datasets to create terms, as seen in Chelvachandran and Jahankhani [2019]. They discussed the keyword analysis of two different open-source datasets. One of the datasets is a set of texts from 15 issues of "Dabiq" and seven of "Rumiyah", both magazine publications in English previously mentioned as ISIS propaganda focused on recruitment. The second dataset is 17000 tweets from more than 100 pro-ISIS accounts from Twitter. The study compared the keywords found in both datasets. The importance of a word was based on the word's frequency, meaning how frequent the terms were defined as significant.

3.4.5 Term-based Method

In Chatzakou et al. [2020], two datasets were used for the experiment with the detection of terrorism and abusive language; one called the "Abusive Dataset" and the "Terrorism Dataset". The "Abusive Dataset" was provided from a previous study, while the other, the "Terrorism Dataset", was created by the researchers and was collected between February 2017 to June 2018. The approach was to collect data by searching for terms related to terrorism (terms in Arabic) in the Twitter API. The set of words was created by Law Enforcement and domain experts. A combination of terms with logical operations was used in López-Sáncez et al. [2018]. Data to find radical far-right Spanish Twitter users was collected through a "logical" term search. The stream of tweets from Twitter, together with terms and logical operators as "AND" and "OR", managed to find Twitter users. A human expert then analysed the users in the last part to assure quality. This data was then used to measure the degrees of radicalisation and a proposed equation was used to detect far-right users.

3.5 Root of Radicalization

In Fernandez et al. [2018], they used a computation approach to detect and predict the radicalisation influence on users based on the social science model "Root of radicalisation". The "Root of radicalisation" model originates from the field of social science. The radicalisation process is divided into three different processes; **Micro**, **Meso**, and **Macro**. The measurement of the effects of micro, meso, and macro is performed as the entire timeline of the user, the subset of shared posts, and the set of URL-links contained in the different posts. The theory is divided into three parts referred to as roots:

Micro: Meaning the roots of radicalisation on an individual level.

Meso: Meaning the roots of radicalisation at group level. The action of meeting like-minded people sharing the same thoughts and being influenced by communities.

Macro: Meaning the roots of radicalisation at society level. Political policies or decisions by the government create a feeling of separation, creating room for radical groups to grow.

The theory showed promising results for detecting radical users and good result prediction by combining them with collaborative filtering.

Chapter 4

Experiments

This chapter will explain the ideas behind each experiment performed in this Master's thesis and how they cover their related research question. Section 4.1 will present each RQs and explain the thought behind each experiment and its purpose. In Section 4.2 the execution of the experiments will be presented.

4.1 Experiment Plan

4.1.1 Research Question 1

RQ1: *Is there a method for detecting far-right users within another domain of radicalisation detection?*

Based on the finding in the preliminary studies, the method used in the paper [Fernandez et al., 2018] called "Understanding the Roots of Radicalisation on Twitter" was discovered. It showed promising results in detecting and predicting Islamic radicalisation and was selected as the primary method used in this master thesis. The decision to use this method will be explained in further detail in Section 6.1.1. The method requires a dataset where users are labeled as radical or not. The method uses entire user profiles containing all the posts made by the user. In the case of Fernandez et al. [2018], the users stem from the dataset *How-ISIS-Uses-Twitter* (see Section 3.4.2) from Twitter. The need for a dataset is the idea behind RQ2. Furthermore, a set with terms defined as radical terms is needed for the method. This need is the reason behind RQ3. Moreover, given the time of publication, modifications, and implementation of newer approaches could improve the results. The publication of Fernandez et al. [2018] was in 2018, and many improvements have been created since that. This

is the fundamental idea behind RQ4.

	P	R	F1	P	R	F1	AvgF1
J48	0.862	0.853	0.857	0.870	0.879	0.874	0.866
N Bayes	0.904	0.895	0.899	0.907	0.916	0.912	**0.906**
Log R	0.901	0.863	0.882	0.883	0.916	0.899	0.891

Table 4.1: The classification results after 10-fold in Fernandez et al. [2018]

Experiment 1: *Does a tailored implementation of the selected method from RQ1 show similar results?*

Experiment 1 is designed to test an implementation of the proposed method from Fernandez et al. [2018] following each step. The method calculates the user's three levels of radicalisation, called micro, meso, and macro. Micro is the radicalisation at the individual level. The users can represent this step as frustration, perception of injustice, or symbolic threats [Fernandez et al., 2018, p.3]. The micro is calculated by using the individual posts published, referred to as the original posts. This means all the posts the user wrote themselves, excluding resharing and retweeting. The original post is used to calculate the micro value, which is the cosine similarity between the original posts and at the radical terms. The similarity is a value between 0 and 1.
Meso is radicalisation at a community or group level. The users represent this step by beginning interacting with people with similar beliefs. This action of interacting is defined as the resharing of the post by other users. The calculations are similar to the one in micro. The only difference is the post used. The post used are the shared post by the user, referred to as the sharing posts.
Macro is the radicalisation on a global level. The user's radicalisation influences governments and societies. This level of radicalisation is excluded due to the complexity of the problem. The idea, as in Fernandez et al. [2018], was to use the URL links in the posts and extract the content from the sites. The problem is that the sites in the posts are from multiple places. To perform a web scraping of the sites would be very time-consuming. Therefore the macro level was excluded.

After performing Experiment 2 and Experiment 3, the needed datasets and terms were created. The experimental plan for Experiment 1 will be to implement it as described in Fernandez et al. [2018]. The results will be compared to the results from the study, as shown in Table 4.1, where the focus will be on the average F1 score called *AvgF1*. The reason is that the F1 score considers both the recall and precision, providing a broader overview of the classifier's performance. The hoped outcome of the experiment is to achieve similar results as in Fernandez

et al. [2018], meaning 0.901 as shown in Table 4.1. If the results of the implementation are equally promising, this could imply that the method is independent of the radicalisation type. The results achieved in Fernandez et al. [2018] can be seen in Figure 4.1, where the micro and the meso values were used separately. Each value was used to train the classifier individually and got their Precision (P), Recall (R), and F1-score (F1) calculated. The *AvgF1* is the average of both micro's and meso's F1-score. Few assumptions about the implementation were made due to missing details in explanations in Fernandez et al. [2018] and will be clarified in section 4.2.1. Additionally, the implementation in Fernandez et al. [2018] does not clarify the type of n-gram used, and the reimplementation of Experiment 1 only uses uni-grams.

4.1.2 Research Question 2

RQ2: *Does a dataset suitable for detecting far-right users exist? If not, how can it be created?*

For the thesis, two different datasets are needed; A radical and a non-radical dataset. Most research in recent years has focused on the classification of publications or posts. They tried to classify if a post contained extreme content or hateful content. The problem with the approaches is that most datasets available are a set of many posts labeled as containing or not containing extreme content. Few are focusing on the entire user profile, containing all the posts of a radical user. This thesis explores the possibility of using the entire user profile and changing the scope from post-level to profile-level (see Figure 4.1). Hence, the dataset structure needed would be a set of radical user-profiles and non-radical user profiles.

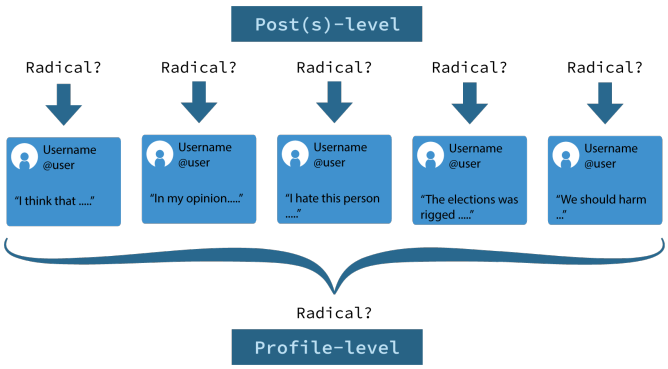

Figure 4.1: Post and Profile level of data

Experiment 2: *Can a radical dataset for detection of far-right/alt-right users be created?*

The plan for Experiment 2 is to use the social media platform Gab (see Section 2.1.3) as a source of radical users. The method uses a post from 2021 by a politically active person from the alt-right. The post encouraged users to comment in the comment section to find "new friends". The idea was for users to comment if they had been removed from Twitter or any mainstream social media platforms and find users on Gab in the same situation. The plan of the experiment was to extract usernames in the post's comment section to create a list of "banned" users. The following step is to assure that the users are actually suspended by checking it through the Twitter API. This step is based on the study in Wang et al. [2016] where it was found that even with the possibility to change or use a new username on other platforms, users tend to fall back to a similar or the same name. The hypothesis was that users that self-reported on Gab to have been banned would have had the same name on Twitter. Users that got a "Suspended" message back from the Twitter API, meaning the "Error 63", was added to a verified radical list. The last step would be to scrape and extract all the posts published by the user on Gab. The set of verified radical users from Gab will be named "The Radical Dataset", where users from this dataset are labeled as radical users. This labeling is because Gab has been defined as an alt-right/radical site multiple times Goodwin [2021]. Further, Gab users with the same username banned on Twitter assure some form of higher radicalisation contrary to regular users.

Experiment 3: *Can a non-radical dataset for detection of far-right/alt-right users be created?*

The plan for Experiment 3 is to use Twitter as a source of non-radical users. Using the Twitter API, the idea is to collect the stream of live tweets created on the platform. One of the Twitter API endpoints, Stream[1], can be used for the task. It returns 1% of all Tweets in real-time. The API will be used for a specific time span, and all the posts will be stored. Each post will be controlled to have the language attribute set to "en", meaning the post is in English. The username of the post's creator will be extracted and stored in a list. The list of usernames is now defined as English-speaking active non-radical users of Twitter. Like the radical users, the users are defined as a non-radical duo to their data origin; Twitter. The users and their posts will be extracted to create the "Non-radical dataset". The users of this dataset are called regular or non-radical

[1] `https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/api-reference/get-tweets-sample-stream`

users.

The purpose of using the live stream is to ensure that the user is currently an English-speaking active user of the platform. The reason behind using only English-speaking users is that English-speaking users dominate Gab.

4.1.3 Research Question 3

RQ3: *What are the advantages and disadvantages of using an unsupervised method to find radical terms?*

The task of making a dictionary with radical terms is a challenging one. As seen in other studies, domain experts are usually used to create the terms. Terms from organisations that focus on radicalisation can also be used. However, it seems to commonly be either another type of radicalisation or not publicly available. In the research found in Chapter 3, no paper had a method or public dataset which could be used. The research on far-right radicalisation is small. Duo to the missing dictionary of radical terms, the task was to create one for the implementations in the Master's thesis.

Experiment 4: *Can an unsupervised approach be created to extract far-right radical terms?*

An experimental approach was suggested in this thesis for creating far-right radical terms. The theory behind the idea will be presented, followed by the implementation. The essence of the idea is to extract radical terms from manifestos of far-right terrorists.

Theory

Let's assume it exists a theoretical point in the latent space representing meaning of words, $\vec{P}_R$, that is defined as the meaning of far-right radicalism. For every word in the English language, $w \in L_{Eng}$, there exists a vector space position where the word is represented, $\vec{w} \in \vec{L}_{Eng}$. With the assumption that words can "contain" a form of meaning and can be represented as a vector position, $\vec{w}$. We assume that all words and $\vec{P}_R$ exist in the same universe of "meaning".

$$\{\vec{P}_R \cup \vec{L}_{Eng}\} \subset \mathcal{U}$$

Give that all words have a position in this vector-space and $\vec{P}_R$ shares the same universe, each word will have a distance and similarity (see Section 2.6) between

itself and $\vec{P}_R$. Each similarity will represent how "radical" every word in the English language is. The task is to create an optimal dictionary of terms containing radical words, a finite set $W_{\vec{P}_R}$ with the "perfect" distance(k) from the $\vec{P}_R$ is the goal.

$$W_{\vec{P}_R} = \{w | w \in \vec{L}_{Eng} \wedge k \leq \frac{\vec{w} \cdot \vec{P}_R}{||\vec{w}|| \cdot ||\vec{P}_R||} \leq 1.0 \}$$

If the exact position of $\vec{P}_R$ was known, a set could be created by retrieving the N closest words in $\vec{w} \in \vec{M}$. The reality is that this position is hard to find and only theoretical, as explained at the start. Additionally to this problem of the missing $\vec{P}_R$, the decision on the perfect k-value would be challenging to decide. If the k-value is too large, too many words/terms would be in the set and potentially be too vague relative to representing far-right radicalism. If the k-value is too small, too few words would be in the set and potential exclude relevant terms.

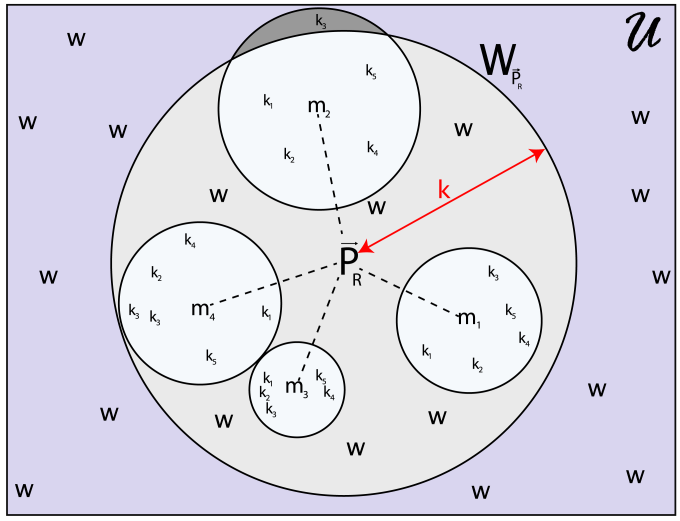

Figure 4.2: An illustration on the theory of optimal retrieval of far-right terms

Given that we are currently talking about the vector-space of meaning, a given position would represent some form of meaning. Positions close to each other share meaning, while positions far from do not. Instead of trying to find the exact position of $\vec{P}_R$ and the optimal k-value, an approach could be to try to come as close to this position as possible. Let's say it exists a set, $\vec{M}$, of vectorised

documents that exist in the same universe, $\mathcal{U}$.

$$\{\vec{m} \in \vec{M}\} \subset \mathcal{U}$$

$$M = \{m_1, m_2, m_3, m_4\}$$

The set of $\vec{M}$ consist of documents defined in the real world as far-right containing literature. The assumption is that the position of each element in $\vec{M}$ would, in theory, be close to $\vec{P}_R$. By now having more defined positions than $\vec{P}_R$, the process is now to find best terms given the position of $\vec{m}$ in $\vec{M}$. The thought is that the $\vec{m}$ will have a chance of overlapping with the terms in the given point, potentially contain sufficiently good radical terms as seen in Figure 4.2. The theory is also supported with the rule of transitivity. Given that $W_{\vec{P}_R}$ is close to point $\vec{P}_R$, and the set $\vec{M}$ is close to $\vec{P}_R$, them the assumption is that M is close to $W_{\vec{P}_R}$. Additionally, the k-value problem is now converted to multiple instances of finding the best k-value/n-value of terms needed to extract from each $m \in M$. Each set with the best terms defining m_n is declared as K_n where the n is the same n from $m_n \in M$.

$$K_n = \{w_1, w_2, ...w_x\} \in m_n$$

$$W_{\vec{P}_R} \approx \{w | w \in L_{Eng} \wedge w \in \{K_1, K_2, K_3, K_4\}\}$$

Now a dictionary/set with radical terms can be created based on finding the best keywords/terms for each $m \in M$. Each set of keywords best representing a $m \in M$ can be added and construct the set of $W_{\vec{P}_R}$ dictionary and can be more securely defined as "radical" terms. The implementation is presented in the following section.

Plan

The plan for Experiment 4 is to take all the manifestos and convert them to a text file. The manifestos will not incluced pages as the table of content or pages with sources. The idea is to remove text not created by the terrorist. The entire text will be used without preprocessing since KeyBERT can use the entire text without preprocessing. The results from KeyBERT are structured in a list of tuples, where each word and similarity score are together. The results sorted and be saved in each corresponding CSV file.

4.1.4 Research Question 4

RQ4: *How can the existing approach for detecting radicalisation be improved?*

The goal of the research question and experiment aims to contribute to the field and improve the method used. The selected method stemmed from 2018, and many discoveries have been made in NLP. Three different modifications were suggested based on the thoughts during implementation in Experiment 1.

Experiment 5: *What can be improved of the method selected in RQ1?*

The definition of improvement combines the baseline results from Experiment 1 and the results found in Fernandez et al. [2018]. The metrics used to evaluate the improvement are precision, recall, and F1 score, where the F1 score is in focus. This is because F1 is a metric that combines precision and recall. Based on the implementation experience in Experiment 1, the are three different ideas for enhancing the system; adding more radical terms, method of vectorisation of micro/meso, and the weighting of radical terms. Each improvement will be implemented and compared to the baseline model created in Experiment 1.

M1: Modification to Number of Radical Terms

The first improvement is extending the terms' size from only 305 terms to including all terms found in Experiment 4. Including more terms could be beneficial for the system to better calculate the radicalisation of each user. There can be critical words excluded by limiting the radical terms only to 305. The experiment will include all the terms found in Experiment 4, meaning 2764 terms from all four manifestos. The modification will change the radical vector, which approximates the radicalisation score in the user's micro and meso. The radical vector is a vector with the length of the number of unique words found in the dataset. However, if the word is found in the radical dictionary (from Experiment 4), it is represented with one. Otherwise, the value is 0. In other words, the radical vector is a vector where each word is represented only if it exists in the radical dictionary.

M2: Modified Vectorisation

The second modification experiments with two minor changes to the user's micro and meso vectorisation. As the implementation in Experiment 1 and Fernandez et al. [2018], each vector for the micro and meso was created by considering every word in the dataset, called the set of W_p.

$$\overrightarrow{Micro}_u = (p_1, p_2, ...p_n), p_i \in P_{uo}$$

$$\overrightarrow{Meso}_u = (p_1, p_2, ...p_m), p_j \in P_{ur}$$

$$V\overrightarrow{micro}_u = (w_1, w_2, ...w_n), w_i \in P_{uo} \wedge w_i \in W_p$$

$$V\overrightarrow{meso}_u = (w_1, w_2, ...w_m), w_j \in P_{ur} \wedge w_j \in W_p$$

Challenges to this method are that the size of W_p will increase with time when more users are added, and creating huge sparse matrixes. The suggested modification is only to use terms from the radical dictionary since the word in the dictionary best represents radicalisation in theory. In other words, means changing the W_p in $V\overrightarrow{micro}_u$ and $V\overrightarrow{meso}_u$ with L, the radical dictionary from Experiment 4.

$$\overrightarrow{Micro}_u = (p_1, p_2, ...p_n), p_i \in P_{uo}$$

$$\overrightarrow{Meso}_u = (p_1, p_2, ...p_m), p_j \in P_{ur}$$

$$V\overrightarrow{micro}_u = (w_1, w_2, ...w_n), w_i \in P_{uo} \wedge w_i \in L$$

$$V\overrightarrow{meso}_u = (w_1, w_2, ...w_m), w_j \in P_{ur} \wedge w_j \in L$$

The change changes the length of the micro and the meso vector and the length of $\vec{L}$, where it becomes a vector with only "1" with the size of L.

$$\vec{L} = (1.0, 1.0, ...1.0)$$

$$|\vec{L}| = |L|$$

The radical vector, $\vec{L}$, is used to calculate the radical score for the user's micro and meso vector.

The second minor change is to the vector representation for a word in the micro and meso vectors. As implemented in Experiment 1 and Fernandez et al. [2018], the value in the vector is created by using the number of occurrences of a term in the user's posts, divided by the number of posts. When calculating the micro vector representation, the following function is used to vectorise each term:

$$val(w_i) = \frac{freq(w_i)}{|P_{or}|}$$

When creating the meso vector, the same function is used. The only modification instead of using the original posts, P_{or}, the shared posts, P_{ur}, is used.

$$val(w_i) = \frac{freq(w_i)}{|P_{ur}|}$$

The implementation used in Experiment 1 and Fernandez et al. [2018] can be seen as a Bag-of-word method with a normalization. The approach counts the occurrences throughout the posts and uses it to calculate the micro/meso vector. The main reason to experiment with this element is that it can occasionally become bigger than the value one. If the occurrence of a term is higher than the number of posts, the value will surpass the value 1, where the limit is unknown. Without a know limitation to each weight of these terms, the calculation of radicalisation of both the micro and the meso radical score can be wrongfully calculated. As it was assumed, the $\vec{L}$ is a vector where the value one is assigned if it is a radical term and zero if not. In theory, by calculating the cosine similarity between the micro/meso vector and $\vec{L}$, the weights of two terms, where one is 0.8 and the other is 1.2, will be equally similar. One can assume that the higher the value of a radical term, the more radical a person is. The proposed modification tries to tackle the problem.

The modification changes the bag-of-words approach with a one-hot-encoding approach. Instead of using the word frequency for each word in W_p, the number of posts counting in the user's posts is used. At most, the numerator will be the same as the denominator if it appears in every post, giving the terms the value of 1. This creates an upper limit where the value at most will be one, making the calculation of the radicalisation score more accurate. The change gives us two new formulas for vectorising the micro/meso vector:

$$val(w_i) = \frac{|\{p|p \in P_{or} \wedge w_i \in p\}|}{|P_{or}|}$$

$$val(w_i) = \frac{|\{p|p \in P_{ur} \wedge w_i \in p\}|}{|P_{ur}|}$$

M3: Adding Two Metrics

The third modification will experiment with adding two metrics. The first metric is a profanity ratio, and the second is the average length of the posts. Research has shown that both features/metrics deliver promising results in detecting radicalisation, as in Lara-Cabrera et al. [2019] and Nouh et al. [2019]. The two features added are metrics for detecting frustration and one on introversion. The belief is that adding more data, especially meta-data data, can improve and be better than the performance in Experiment 1 and Fernandez et al. [2018]. Meta-data means information about the other data. Both metrics will also be divided into micro and meso, meaning there will be created a two profanity ratio and two post averages in correlation to either the original posts or shared posts. The two additional metrics will hopefully, together with the micro and meso value,

help the classifier detect far-right radicalisation better. Experiment 5 will include the metrics and compare the results found in Experiment 1 and Fernandez et al. [2018]. Furthermore, the experiment is a continuation of modification M2, so only the 305 radical terms are vectorised when calculating the micro and meso vectors. The findings in Lara-Cabrera et al. [2019] highly inspire both metrics.

4.2 Execution

The experiments are inspired by the approach presented in Fernandez et al. [2018]. As mentioned in Section 3.5, the method is used to detect and predict Islamic radicalisation. The adjustment of the original method to far-right radicalisation detection created the task of creating data. The task was finding both datasets and creating far-right terms. Consequently, the order of experiments was changed. The datasets and the radical terms needed to be created before the implementation in Experiments 1 and 5. The execution order became first Experiments 2, Experiment 3, and Experiment 4, and then Experiment 1 and Experiment 5.

4.2.1 Experiment 1

Experiment 1: *Does a tailored implementation of the selected method from RQ1 show similar results?*

The implementation in Fernandez et al. [2018] utilised two datasets originating from Twitter containing 112 radical users and the 95725 regular users. The radical dataset originated was the famous "How Isis Uses Twitter" dataset. In contrast, the non-radical dataset is a public dataset called "isis-related-tweets", where users were extracted with ISIS-related keywords. One hundred twelve users were randomly selected from the public dataset that was currently active on Twitter. The two annotators, authors of the paper, verified 40 of the 112 users by manually checking if they showed no sign of supporting ISIS. This became the non-radical dataset. A dictionary containing radical terms was constructed by combining multiple public dictionaries created by institutions working on radicalisation. The radical dataset is created in Experiment 2, while the non-radical is created in Experiment 3. The radical terms of far-right radicalisation are created in Experiment 4. By having all the different data needed, Experiment 1 can fully implement the approach found in Fernandez et al. [2018].

Dataset and Radical Terms

One hundred twelve users were randomly selected on each dataset, creating a new dataset. The datasets are the radical dataset and the non-radical dataset. The decision to use 112 each stems from the implementation in Fernandez et al. [2018] to be true to the original implementation. This assures that the results can morse justifiably be compared. Additionally, a set of radical terms from Experiment 4 are used to create the radical dictionary. The first 305 terms from the dictionary with the highest similarity are used. The specific number of 305 terms stems from the implementation in Fernandez et al. [2018].

User's post to vectors: Micro and Meso

Each user got their micro and meso vector calculated by utilising the entire user profile, meaning all the posts created by the user. The micro and meso level of radicalisation was included in the implementation during the experiment. In contrast, the macro-level was excluded due to the difficulty of extracting information from the URLs and external sites. The following two sections explain how the user posts were vectorised to micro and meso vectors to calculate the values.

Micro

The micro vector represents the radicalisation on the individual level. As in Fernandez et al. [2018], $\vec{Micro}_u$ represent the entire timeline for a user u on a platform, P_u. P_{uo} represents the post created by the user u, the original posts, and is a subset of all posts found on P_u.

$$\vec{Micro}_u = (p_1, p_2, ...p_n), p_i \in P_{uo}$$

The posts from $\vec{Micro}_u$ are used to create $\vec{Vmicro}_u$ for each user u, where it consist of all words found in $\vec{Micro}_u$ and W_p. W_p represent all unique terms found in the entire dataset.

$$\vec{Vmicro}_u = (w_1, w_2, ...w_n), w_i \in P_{uo} \wedge w_i \in W_p$$

$$val(w_i) = \frac{freq(w_i)}{|P_{uo}|}$$

The final vector, $\vec{Umicro}_u$, representing each user u is created by assigning weights to each word in $\vec{Vmicro}_u$ with the $val(w_i)$. The weighted vector, $\vec{Umicro}_u$ is used to represent each user and is the vector used to calculate the radicalisation at the micro level. Each weight is created by taking the frequency of the word and dividing it by the number of original posts created by the user u, excluding

shared posts. The division by the original post number is used to normalise the weight. The final micro vector for user u is:

$$U\vec{micro}_u = (val(w_1), val(w_2), ...val(w_n)), w_n \in W_p$$

Meso

The meso vector represents the radicalisation at the group level and is created by only using the post shared by the user, excluding original posts.
As in Fernandez et al. [2018], $\vec{Meso}_u$ represent the entire timeline for a user u on a platform, P_u. P_{ur} means the post reposted or shared by users u, the *reposts*, and is a subset of all posts found on P_u.

$$\vec{Meso}_u = (p_1, p_2, ...p_m), p_j \in P_{ur}$$

The shared posts from $\vec{Meso}_u$ are used to create $\vec{Vmeso}_u$ for each user u, where it consist of all words found in $\vec{Meso}_u$ and W_p. W_p is the same as when calculating the micro vector, where it represents all unique terms found in the entire dataset.

$$\vec{Vmeso}_u = (w_1, w_2, ...w_m), w_j \in P_{ur} \wedge w_j \in W_p$$

$$val(w_i) = \frac{freq(w_i)}{|P_{ur}|}$$

Terms are weighted with a similar $value(w_i)$ as in micro, but instead are divided/normalised with the number of total shared posts, meaning $|P_{ur}|$. The user u meso vector is therefore represented as:

$$U\vec{meso}_u = (val(w_1), val(w_2), ...val(w_n)), w_n \in W_p$$

Calculation of radicalisation

As seen in Fernandez et al. [2018], the dictionary L containing the radical terms is vectorised to create $\vec{L}$. The vector is used with the cosine similarity (see Section 2.3.8) on the user's micro and meso vector to calculate the radicalisation score/value. The desire is to find the influence, meaning similarity between the user's vectors and the radical vector $\vec{L}$ to estimate radicalisation influence. The idea is the higher the influence/similarity, the more radical a person is in the given level of radicalication. The creation of $\vec{L}$ is not explicitly explained, so assumptions were made. The vector $\vec{L}$ was the same length as the vocabulary of the dataset, W_p. Each position in the vector represents a unique word. If the current position, meaning word, exists in the radical dictionary, the weight in the

vector position is given one. The value is set to zero if it is a word that does not exist in the radical dictionary.

$$radicalValue(w_n) = \begin{cases} 1.0, & \text{if } w_n \in L, \\ 0.0, & \text{otherwise.} \end{cases} \tag{4.1}$$

The binary "weighting" and creation of $\vec{L}$ is a modified version of $val(w_n)$. The new version was called $radicalValue(w_i)$ was used to construct $\vec{L}$.

$$\vec{L} = (radicalValue(w_1), radicalValue(w_2), ...radicalValue(w_i)), w_i \in W_p$$

Now by having the $\vec{L}$, each user in the dataset with 224 users got their micro and meso vector compared to the $\vec{L}$ to create a micro and meso radicalisation value. The value returned after performing the cosine similarity was stored and used to train and test the different machine learning classifiers.

Machine learning

Three machine learning classifiers were trained and evaluated with the metrics precision, recall and F1; a Decision Tree-, a Naïve Bayes-, and a Logistic Regression classifier. Given that there were no parameters specifications in Fernandez et al. [2018], almost all the values of each machine learning classifier were only based on the default values provided by the library, Scikit-learn[2]. Only the decision tree got its parameters changed. Contrary to the implementation in Fernandez et al. [2018], the J48 classifier was replaced with a Desition Tree. J48 is a decision three implemented but implemented in a software called WEKA, where it is based on the C4.5 algorithm and utilises information gain to decide on the feature to use [Mashiloane and Mchunu, 2013, p.544]. Due to not having access to the WEKA software, the implementation changed to use the decision tree classifier from Scikit Learn; since its implementation uses the tree algorithm CART, which shared many similarities to C4.5[3]. The classifier's criterion was also set to "entropy" since information gain uses entropy. This is because J48 uses those parameters.

All parameters can be found in Appendix C, where the different parameters are shown and explained in a short explanation.
The training and testing were performed in a 10-fold testing. The dataset for training is divided into ten equally sized parties and "rotates", in which one is

[2] `https://scikit-learn.org/stable/index.html`
[3] `https://scikit-learn.org/stable/modules/tree.html`[Accessed 10.06.2022]

tested to estimate the performance of the classifiers better. The scores are averaged by the number of folds to get an average value on precision, recall, and F1 for each classifier. The results can be seen in Table 5.1.

4.2.2 Experiment 2

Experiment 2: *Can a radical dataset for detection of far-right/alt-right users be created?*

The data used to construct the dataset originates from the social media platform Gab. As mentioned is section 2.1.3, the platform is seen as hosting multiple radical users, mainly on the right side of the political spectrum. Additionally, Twitter is seen to be a preferred choice for research as seen Section 3.4.2 in due to its accessibility and well documentation of its API.

Step 1: Origin of radical users/usernames
On the 7th of December of 2021, a public figure with relations to alt-right ideology published a comment on the platform Gab. The post talked about the strict rules of the similar platform Twitter and encouraged if you, as the reader, had been banned/suspended to find new friends on Gab. The idea was that users in the same situation could meet each other in the comment section by commenting "I need friends" and adding each other. On the 1st of February 2022, the published post had reached approximated 1200 comments on its commenting section.

Step 2: Scraping and verification of radical users
A Python script using the frameworks Requests[4] was used to scrape the specific site with the post and store it as a static HTML file. The library Beauitfulsoup[5] was used on the HTML file to scrape the users in the commenting section and their unique usernames. The result is 1200 usernames from users claiming to be banned/removed from Twitter. To further verified is the users' claims were truthful, each username was verified through the Twitter API.

Step 3: Verification of the users and user scraping
A study in Wang et al. [2016] found that even with the possibility to change or use a new username on other platforms, users tend to fall back to a similar or the same name. The hypothesis was that users that self-reported on Gab to have been banned would have had the same name on Twitter. If a user were removed, the API would not find the user and return "User

[4] `https://docs.python-requests.org/en/latest/`
[5] `https://www.crummy.com/software/BeautifulSoup/bs4/doc/`

not found", making it impossible to distinguish between if the user has a different username or were banned. With a similar approach as in Ali et al. [2021], users got assumed to be suspended only when the error code "403 [63 - User has been suspended.]" was returned by the Twitter API. Out of the 1200 users, 313 users returned with the suspension error code and were therefore classified as radicals. Using the library Garc[6], an API wrapper for retrieving information from Gab, the user's entire profile was scraped for posts and saved as a JSON file. A complete set of each radical user's profile posts became the dataset known as the "radical" dataset. The extraction was performed on the 3rd of February 2022.

4.2.3 Experiment 3

Experiment 3: *Can a non-radical dataset for detection of far-right/alt-right users be created?*

The data source for the creation of the non-radical dataset became Twitter. As mentioned in Kor-Sins [2021], Gab has, during a period of high moderation in mainstream platforms, promoted itself as a censorship-free platform. An unintentional outcome of the regulation has created an exodus of alt-right users to the Gab platforms from platforms like Twitter [Kor-Sins, 2021]. Therefore, the assumption is that extracting data from Gab And Twitter will not create much of a difference since they share the same target group of users.

Step 1: Get random users from Twitter
The gathering of data was done by writing a Python script utilising the library Tweepy[7] and using the Stream API[8] from Twitter. The API returns a live stream of user samples on the platform currently publishing tweets, assuring the users are current Twitter users. The username was extracted and saved on a list if the post's language was declared as English. The extraction of usernames was performed on the 27th of April 2022.

Step 2: Extract the user profiles
The usernames were retrieved with the User API to get the user's id and extract all their posts. The posts were stored as a JSON file together with post information as the time of publication, the content and if they were an "original" post or retweet. For better dataset quality, the API call was limited to return only posts before the 3rd of February 2022 to assure that

[6] `https://github.com/ChrisStevens/garc`
[7] `https://docs.tweepy.org/en/stable/`
[8] `https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction`

both datasets were from the same period. Due to complications with the API and other limitations, the total number of "normal" users extracted became 213 users. It is explained more in details in Section 5.3

4.2.4 Experiment 4

Experiment 4: *Can an unsupervised approach be created to extract far-right radical terms?*

Further presented in detail in Section 4.2.4, a set of manifestos published online by alt-right/far-right extremists published before a terror event are used in Experiment 4. The task of finding a radical dictionary is by converting the task further to an already solved known problem in NLP, keyword extraction of documents. Keyword extraction is the task of extracting the best word or n-grams to represent a document best. Today, multiple approaches have been developed over time, such as RAKE, TF-IDF, Yale, and KeyBERT[Giarelis et al., 2021]. As discussed in Section 2.5.3, based on a comparative study in Giarelis et al. [2021], the decision fell on utilising the KeyBERT. This is because it has shown to be effective when used on longer text and the manifestos are long (4-1500 pages). KeyBERT process of finding the keyword is explained in Section 2.5.3. In summary, KeyBERT vectorised the documents and, based on the vector, finds the best terms/candidates to represent the documents. The vectors of the document and terms are from the embedding from BERT (see section 2.3.7) The following section presents each of the manifestos used for the creation of the radical dictionary. All terrorists have been declared as right-wing terrorists, from the far-right and alt-right.

Events of Terror and Manifestos

Each of the following events is a terror attack executed by far-right extremism. As mentioned in Section 1.3.2, the belief that violence is a reasonable means to obtain their ideas is the main difference between a radical and an extremist. Since each individual performed a terror attack costing numerous lives, it is safe to assume they share this view on violence. The shared nominator for the event is that they came with a publication of a manifesto explaining and communicating the perpetrator's beliefs. There exist more events where a terror attack and manifesto came together, but the decision fell on only using four for simplicity. Based on the manifestos presented in Ware [2020], the manifestos were selected due to numerous cross-references between the manifestos and also to cover a more extensive geographical area.

- **Manifest 1**: On Friday at 15:25, the 22nd of July 2011, a 950 kg bomb went off in the middle of Oslo. Eight people lost their lives on the spot, and ten others got life-threatening damages. The explosion created enormous damage to buildings and placed multiple people in life-threatening danger. But it was only the start of the attack. A few hours later, the emergency lines got distressed calls from numerous on an island not far from Oslo, Utøya, where it was reported that "*a man in police uniform shooting in the island, many people are down*". Hours before the event, a 1500 pages compendium/manifest with the title "*2083 - A European Declaration of Independence*" by the perpetrator under the pseudonym Andrew Berwick was mailed to numerous people. The attacks by Anders Behring Breivik is seen as the worst attack on the Norwegian democracy since WW2. The attack cost the life of 69 people, where the majority were young people [Gjørv, 2012, p.17-37]. The plan was to attack The Norwegian government and the political party Arbeiderpartiet, The Labor Party [Gjørv, 2012, p.255], and he was multiple times observed sharing utterances about his anti-immigration and hate for the government[Gjørv, 2012, p.360].

- **Manifest 2**: On Friday, 15th of match 2019, two mosques were subject to a terrorist attack in Christchurch, New Zealand. While live-streaming on Facebook, he entered the mosques with a firearm and opened fire. The Australian perpetrator wanted to sabotage the lives of Muslim worshippers. A document also appeared on a site called 8chan before the attacks [Ware, 2020, p.3]. The around 70-page manifesto named "*The Great Replacement: Towards a New Society*" was published on forum 8chan and prompted the theory that immigrants are replacing the western population and culture. He also stated that the attack was politically motivated and defined himself as an "eco-fascist" and racist[Brzuszkiewicz, 2020, p.73]. The actions of Brenton Tarrant cost 50 people there life[Brzuszkiewicz, 2020, p.75].

- **Manifest 3**: In El Paso, Texas, on the 3rd of August 2019, a manifest of 4 pages appeared on social media platforms originating from a forum called 8chan . The manifesto was four pages long and mainly focused on hatred against immigrants of Hispanics [Ware, 2020, p.2]. The manifesto began with support for the Christchurch shooters and justified his action as a response to the invasion of Texas by Hispanics. Later, the perpetrator entered a Walmart at Cielo Vista Mall in El Paso and opened fire [Ware, 2020, p.3]. In his manifesto, he expressed his desire to kill as many possible Hispanic people based on his claims that they are replacing native-born Americans[Brzuszkiewicz, 2020, p.75]. The actions of 21-year old Patrick Crusion cost the life of 22 that day[Ware, 2020, p.3].

- **Manifest 4**: The German neo-nazi Stephan Balliet published documents containing with rhetorics similar to Tarrant's manifesto, including theories as to the "great replacement". The theory is based on the concern about the white people being replaced by "non-whites" since the birthrate is higher in the population with immigration backgrounds. The theory also accuses a "secret Zionist" group with political influence behind the "Great replacement" with the plan of eradicating the white race[Ware, 2020, p.5]. Additionally, the manifesto included his plan to attack the synagogue in Halle to kill non-whites and Jews[Brzuszkiewicz, 2020, p.75]. A week after publication, on the 9th of August 2019, Balliet started a live stream on the Twitch and initiated with opened fire outside a synagogue in Halle, Saxony-Anhalt, killing two persons and injuring two more[Brzuszkiewicz, 2020, p.75].

Method

The set of manifestos contains four manifestos and are from different people of different terror-events. This Master's thesis defines all the terrorists as being on the far-right extremists given their actions and motivations expressed in the manifestos(See Section 4.2.4). Each manifesto was found as a PDF file public on the internet and crossed validated by comparing the document from multiple publication sources. The idea behind utilising the manifestos is that, given their recent publication time (2011 - 2019), they could represent the values shared with the far-right radicals better, contrary to the older literature. The first thought was to utilise books originating from historical people with ties to Fascism and Nazism. Still, given the time difference between the recent rise of far-right ideology and the time of publication, the representation would risk to not be sufficient.

- **Step 1: From PDF to Text**
  The manifesto formats varied from docx and PDF, and the first task was to convert the manifestos to the same format, a Text-file. Manifesto 2 was in the format of Docx, while the rest was in the PDF format. Manifesto 2 was first converted to a PDF file, so the converting task became only from PDF to Text. The library PyPDF2[9] was used the extract the text inside the pdf files. Removal of the table of content and the sources list pages on each manifesto was with the intention of only extracting the written words of the terrorist.

- **Step 2: Text to Keywords**
  After converting PDF to text, the process was to apply KeyBERT to it.

[9] `https://pypdf2.readthedocs.io/en/latest/`

The implementation is easy to use and "plug-and-play, since KeyBERT uses BERT, which takes text without text reprocessing. The text files of the manifestos were read with the native Python "read" function and sent through KeyBERT.

```
def get_n_keywords(doc: str, n: int, n_gram : tuple) -> list:
    """
    Get the n most important keywords from a document.
    """
    k = KeyBERT()
    keywords = k.extract_keywords(doc,
                                  keyphrase_ngram_range=n_gram,
                                  stop_words="english",
                                  top_n=n)
    return keywords
```

Listing 4.1: The KeyBERT function

The results from KeyBERT are in a tuplet format where the elements are a term and the corresponding cosine similarity between the term- and document vector. Both were saved in a CSV file, but only the terms were used to create the end dictionary.

4.2.5 Experiment 5

Experiment 5: *What can be improved of the method selected in RQ1?*
The modifications were implemented on the already implementation from Experiment 1. The changes were mainly in the vectorisation function, where both the micro and the meso vector are calculated, and the number of terms retrieved from the radical ordinary.

The first modification, called M1, was implemented by changing the number of radical terms extracted from the dictionary. The system retrieves all the different terms from Experiment 4 and sorts them based on their similarity score. The top 305 highest scoring terms are retrieved and used for the experiment. M1 changed the number by retrieving all the terms from the radical dictionary. The change results in retrieving 2764 terms, which are used for creating the micro and meso vector.

The second modification, M2, experiments with a new approach to calculate the radicalisation in each user. The modification contains two minor changes in which words are in focus when creating the micro/meso vectors and one on the method of "counting" the terms. The micro and the meso vector are calculated using all the unique words found in the dataset, meaning both the radical and non-radical user's profile. Adding more users over time will increase the number

of unique words, making the radicalisation calculation more processing demanding. The set W_p(all unique words in the entire dataset) was replaced with L, the radical dictionary from Experiment 4. To be as faithful to the implementation of Fernandez et al. [2018], 305 terms were extracted from the radical dictionary to be used to calculate the user's micro and meso. The function in the code where vectorisation of the user's original posts/shared posts is created was modified to implement the new method of vectorising. The new method takes in all the users' posts in a two-dimensional array consisting of the post's words in an array. Each post is changed to consist of only "unique" terms in the post. This assures that each term only is "counted" once in all the posts. The function is also changed to iterate through each word found in the 305 terms from the radical dictionary instead of all words in W_p as in Experiment 1. Each iteration takes one word and counts the number of posts containing that word. The term's value is then divided by the number of either the shared post or original post as normalization. The micro and meso values are then used to train and test the three classifiers; Decision tree, Naive Bayes, and Logistical regression classifier.

The third implementation, M3, can be seen as the continuation of M2. The modification gives a set of new metrics to the radicalisation scores. As explained previously, the metrics will be based on each set of posts, meaning the original and shared posts. The division is to create the metrics with their corresponding "micro" or "meso" values. The values calculated are planned to be used with the classifiers when classifying based on the micro/meso value. The profanity is calculated based on the number of words in the post. Each user gets all their posts reformated to an array of words, which now contain each word they have written. The list of profanities used is from Kaggel[10] and is called "Profanities in English - collection"[11]. The list/dataset contains more than contains 1600+ profanities with their different variations. In online sites and forums, users modify the swearwords before publication to avoid censorship by the site. One example is using the word "@ss" instead of "ass", where the at-sign represents the letter "a". The list of all words from the users which exist in the "profanity list" was counted and divided by the number of all words. The results return the percentage of profanities in all their words. Again, this is performed first on the original posts, then on the shared post, creating two profanity metrics. The second metric is calculated by taking all the posts by the users and finding the average length of the posts. All posts are counted in their "pure" textual form without modification, as posted on the site. Then the sum of the length of all the posts is divided by the number of posts. The value returned is the average length in character for each post by the user. As in the profanity metric, the

[10] `https://www.kaggle.com`

[11] `https://www.kaggle.com/datasets/konradb/profanities-in-english-collection`

average is calculated twice for each user. One is for the original post and one for the shared post by the user. Each user gets four new metrics: two profanity metrics and one average value. Each metric will be used together with the micro and meso value to explore if the results are improved in the model.

Chapter 5

Results

This chapter contains the results from the experiments from Chapter 4. Each experiment is presented and contains some analysis and comments. In Experiment 1, the results of the implementation will be presented. Experiment 2 and Experiment 3 will present the dataset and a minor analysis, which will be discussed in the next chapter. Experiment 4 will present the terms found and contain a minor analyses for the next chapter. Experiment 5 will present the different results from the modifications to the implementation in Experiment 1. All experiments will be discussed in Chapter 6.

5.1 Experiment 1

Experiment 1: *Does a tailored implementation of the selected method from RQ1 show similar results?*

The implementation combined all the radical terms from Experiment 4 and sorted them by their corresponding similarity score. The first 305 terms were used to create the radical dictionary. The intention was to retrieve the best representative terms from all the manifestos used in Experiment 4. The radical and non-radical datasets from Experiments 2 and 3 were combined by extracting 112 users from each of them. This created a dataset of 224 users to calculate the radicalisation and train the classifiers. The choice of using 112 users each stems from the implementation in Fernandez et al. [2018].

In Figure 5.1 the 224 users are plotted. The blue dots represent the regular users, and the red dots represent a radical user. The X-axis represents the value of the user's micro radicalisation score, and the Y-axis represents the meso value.

Both values represent the cosine similarity between the user's micro vector or meso vector and the radical vectors. Using the micro and meso scores, each user gets assigned a position in the two dimensions coordinate system as seen in Figure 5.1.

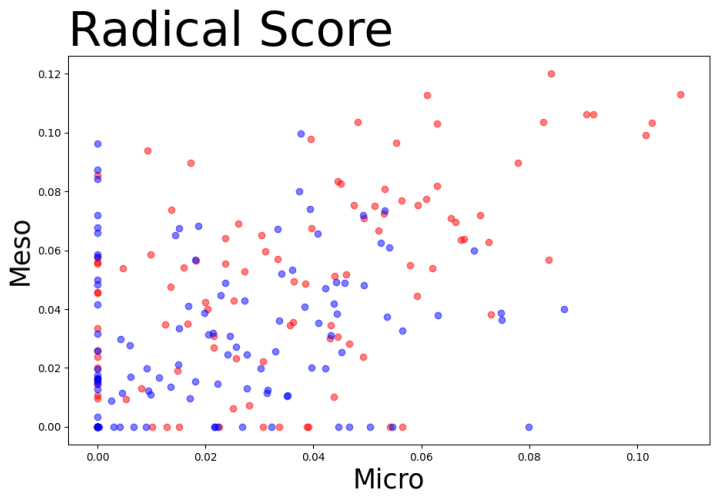

Figure 5.1: A graph showing the 224 users

The result from the implementation is in Table 5.1. The first column is the name of the classifiers. The following precision, recall, and F1 represent the results based on the micro value, while the following precision, recall, and F1 are the result only using the meso value. The last column is the average of the micro's F1 score and meso's F1 score. The bold result represents the highest value achieved. The highest average F1 score achieved is 0.569, which is far less than the 0.901 achieved in Fernandez et al. [2018].

Classifier	Precision	Recall	F1	Precision	Recall	F1	AverageF1
DecisionTreeClassifier	0.534	0.528	0.522	0.503	0.497	0.476	0.499
GaussianNB	0.556	0.552	0.522	0.639	0.637	0.615	**0.569**
LogisticRegression	0.341	0.526	0.378	0.381	0.515	0.36	0.369

Table 5.1: Results from Experiment 1

The test is performed as 10-fold (see Section 2.4.4), meaning the data has been partitioned into ten parts where the partitions for training and testing are changed for each run. The overall results are calculated by averaging all the results, meaning the precision, recall, and F1 score. The inspiration is taken from Fernandez et al. [2018] and used in Experiment 1 and 5.

	Precision	Recall	F1
DecisionTreeClassifier	0.647	0.647	**0.639**
GaussianNB	0.597	0.589	0.577
LogisticRegression	0.636	0.574	0.502

Table 5.2: Results of implementation in Experiment 1. Micro and meso was used during training and testing.

An extra test was implemented in the experiment where the classifiers were trained with both the micro and the meso value. The input to the classifiers was changed from one at a time to including both micro and the meso. The idea was to improve the system's performance. The exact test was also used multiple times in Experiment 5. The results can be seen in Figure 5.2 from this experiment. The value is bold in the F1 column and holds the highest value. The value reached was 0.639.

The results from Experiment 1 are referred to as the baseline results and used to evaluate the improvement accomplished in Experiment 5.

5.2 Experiment 2

Experiment 2: *Can a radical dataset for detection of far-right/alt-right users be created?*

The dataset created in Experiment 2 contains far-right users who were retrieved from Gab and classified as radical users. Each user got all their posts stored and saved together with metadata, such as the date, mentions, hashtags, and unique post-ID given by Gab. Analysis of the dataset, referred to as the radical dataset, shows that it contains 75 788 posts in total by 291 users. The dataset initially contained 313 users when extracted, but after removing users with no posts on their user profiles, the number changed to 291. Of all the posts, almost 61.8% of the posts in the radical dataset were shared. In the original posts by the radical users, there were 151 352 hashtags where 9 233 were unique. However, the shared posts contained 59 130 hashtags in which 9 898 were unique.

	Original	Frequency	Shared	Frequency
1	#democrats	4227	#maga	861
2	#trump2020	3374	#covid	768
3	#joebiden	3362	#democrats	683
4	#cdnpoli	3121	#cdnpoli	675
5	#maga	2998	#trudeau	637
6	#trudeau	2931	#joebiden	597
7	#covid	2436	#trump	594
8	#covid19	2390	#alllivesmatter	545
9	#trump	2162	#backtheblue	543
10	#votetrump	1697	#stopthesteal	531

Table 5.3: Usage of hashtags by radical users

Based on the frequency of hashtags, the ten most popular can be found in Table 5.3 where they are divided between the original posts and the shared posts. Additionally, word clouds representing the hashtags in the original posts by radical users can be seen in Figure 5.2, and hashtags from the shared posts can be found in Figure 5.3. The size of each hashtag represents the relative frequency between all other hashtags, where big means high frequency.

On average, a radical user's profile contains 260 posts, where each post is 374 characters and contains 0.6 hashtags.

5.3 Experiment 3

Experiment 3: *Can a non-radical dataset for detection of far-right/alt-right users be created?*

The non-radical dataset contains 213 users from Twitter, making it 78 fewer than the radical dataset. Complications and challenges with time-outs with the Twitter API made it challenging to extract users' posts, forcing the use of different accounts and VPN to go around the ban. The dataset contains 56 299 posts, where 32 507 were shared/retweeted, giving it a percentage of 57.7%. Analysis of the non-radical dataset showed that posts created by the regular users contained 4101 hashtags, in which 929 were unique. In the shared posts by regular users, it was discovered 9641 hashtags, where 3805 were unique. In Table 5.4,

the ten most frequent hashtags can be seen. The hashtags are divided between the original posts on the right and the shared post on the left. The rest of the hashtags can be viewed in Figure 5.4 and 5.5, where the first is the hashtag from the original post, while the second is the hashtag retrieved from the shared posts in the non-radical dataset.

	Original	Frequency	Shared	Frequency
1	#iheartawards	755	#bitcoin	311
2	#butter	406	#akita	267
3	#bestmusicvideo	406	#beast	195
4	#bestfanarmy	358	#bts	192
5	#btsarmy	357	#bts_butter	163
6	#bts_butter	103	#shib	162
7	#bts	100	#iheartawards	148
8	#footiestories	52	#akitainu	123
9	#gayc	47	#shibarmy	118
10	#bitcoin	43	#1	99

Table 5.4: Usage of hashtags by regular users

On average, a regular user's profile contains 264 posts, where each post is 99 characters and contains 0.25 hashtags.

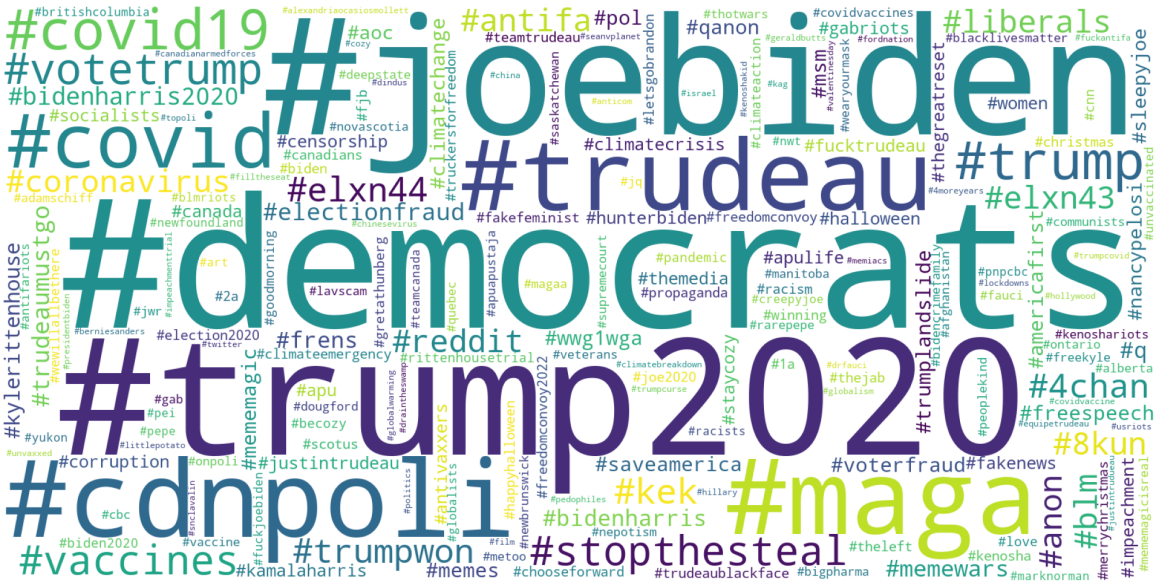

Figure 5.2: Hashtag usage in original posts by radical users.

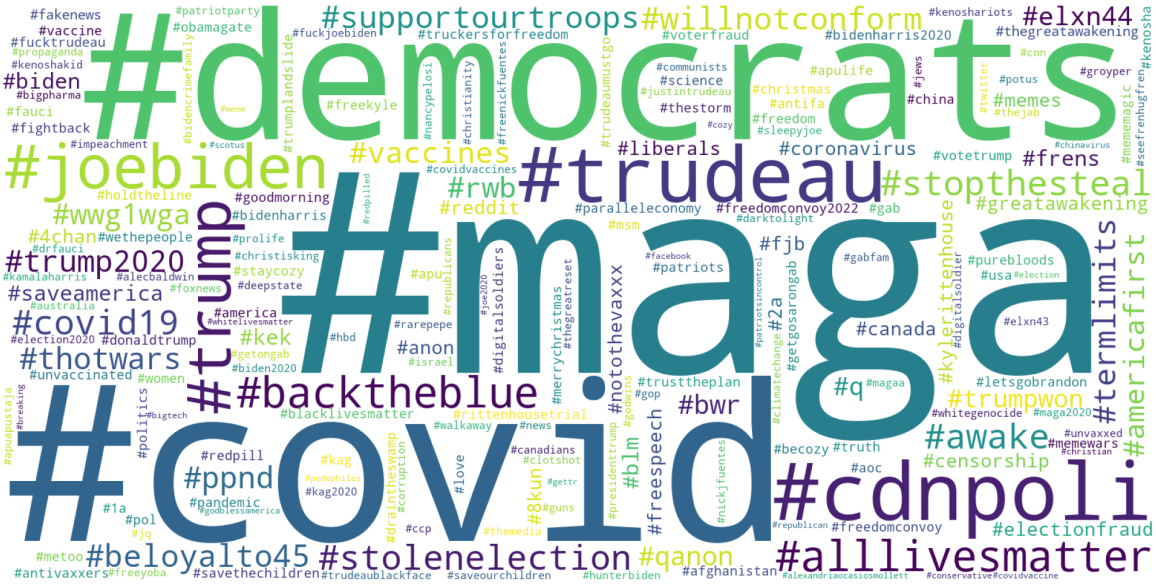

Figure 5.3: Hashtag usage in shared posts by radical users.

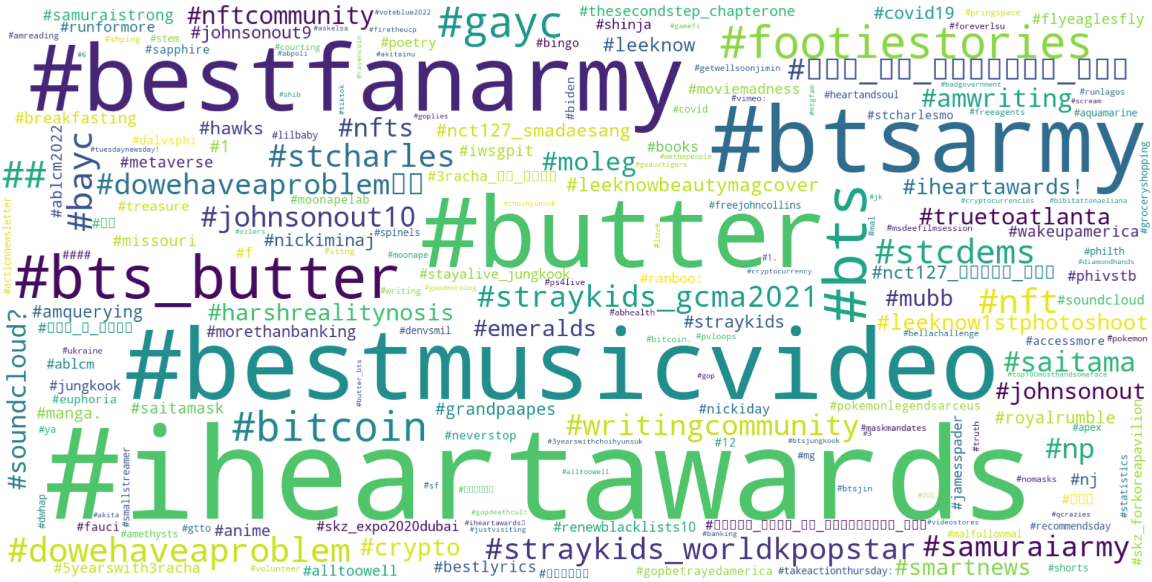

Figure 5.4: Hashtag usage in original posts by normal users.

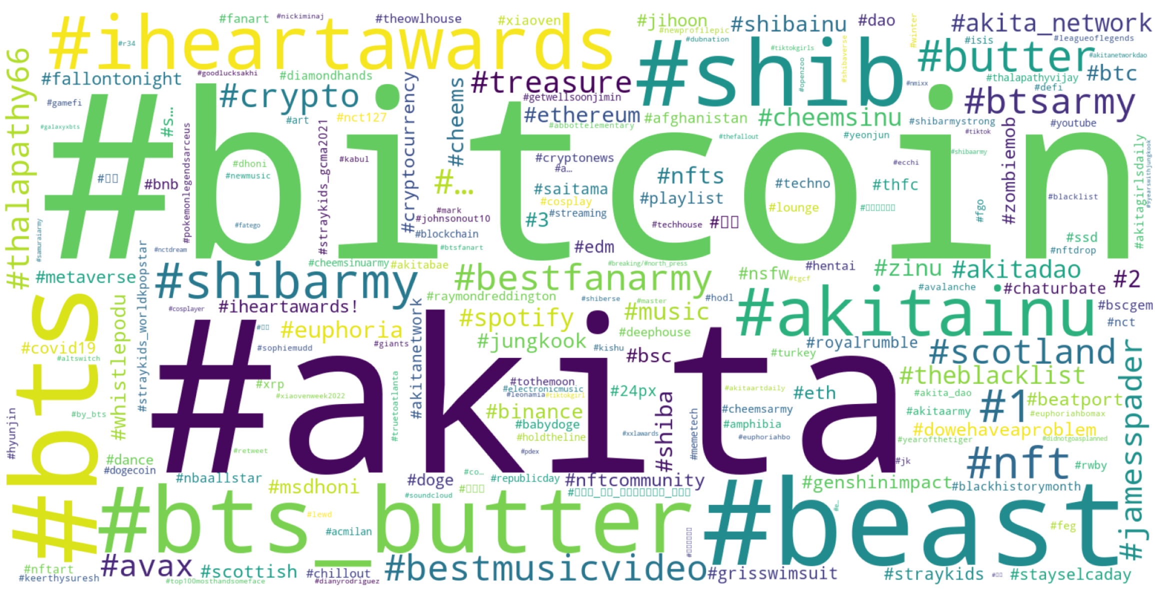

Figure 5.5: Hashtag usage in shared posts by normal users.

5.4 Experiment 4

Experiment 4: *Can an unsupervised approach be created to extract far-right radical terms?*

The results from KeyBERT are presented in Table 5.5 with the ten highest-scoring terms from each manifesto. The terms are sorted based on their similarity score return from KeyBERT. Results from KeyBERT are returned tuplets where a term and the corresponding cosine similarity are together. The similarty score is based on the similarity between the term and document vector.

#	Manifest 1	Manifest 2	Manifest 3	Manifest 4
1	europeanism	exorcism	hispanic	shotgun
2	ideology	manifesto	unrest	ammunition
3	marxism	society	paso	firearm
4	ideologies	exorcist	natives	pistol
5	marxists	blackness	patriotic	gun
6	westernism	inferior	hispanics	ammo
7	repression	manson	threat	carbine
8	counterculturalists	dealing	immigration	shotguns
9	marxist	suffer	deporting	mags
10	ideological	social	immigrant	guns

Table 5.5: KeyBert's extracted keywords

Each of the columns represents the same manifestos presented as listed in Section 4.2.4. For instance, M1 refers to manifesto 1. The limit of KeyBERT was set only to return 1000 keywords/terms, where only Manifesto 1 reached the limit. The different manifestos contributed with:

- **Manifesto 1:** 1000 terms
- **Manifesto 2:** 366 terms
- **Manifesto 3:** 677 terms
- **Manifesto 4:** 721 terms

The variation in the number of keywords is due to the different sizes of the documents. The terms were stored as a CSV file and were used to create the "Radical Terms Dictionary", also known as the radical terms. The radical dictionary is

used in Experiment 1 and 5, where the 305 terms with the highest similarity are used.

5.5 Experiment 5

Experiment 5: *What can be improved of the method selected in RQ1?*

Experiment 5 investigates modifications to the implementation of Experiment 1. The description of how the implementation in Experiment 1 was performed can be found in Section 4.2.1. Only the modifications will be explained in the following sections and will use the same mathematical variable names from Experiment 1.

5.5.1 M1: Increase Number of Terms

The first experimental modification changes the limit on the number of radical terms. As in Fernandez et al. [2018], the terms were selected by combining multiple dictionaries and consisted of 305 terms. The experimental modification changes the number of used terms to include all the terms found in Experiment 4 to improve the precision, recall, and F1 score. Experiment 1 is performed again where the radical dictionary consists of 2764 terms.

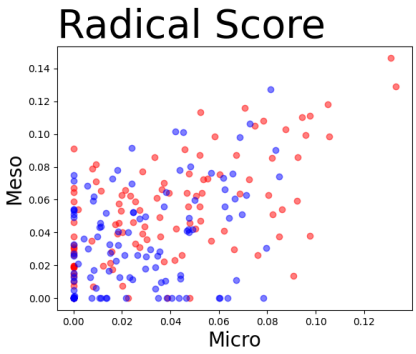

Figure 5.6: Using 305 terms

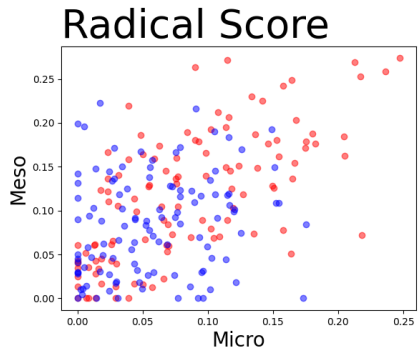

Figure 5.7: Using 2764 terms

In Figure 5.6 and Figure 5.7, each user is plotted based on their radical score in both the micro and the meso value. The X-axis represents the micro value, while the Y-axis represents the meso value.

	Precision	Recall	F1	Precision	Recall	F1	AverageF1
DecisionTreeClassifier	0.550	0.544	0.533	0.451	0.448	0.434	0.484
GaussianNB	0.557	0.547	0.527	0.547	0.532	0.514	0.520
LogisticRegression	0.558	0.563	0.493	0.562	0.568	0.465	0.479

Table 5.6: Results from the first modification (M1)

	Precision	Recall	F1
DecisionTreeClassifier	0.553	0.553	0.541
GaussianNB	0.574	0.571	**0.553**
LogisticRegression	0.559	0.552	0.514

Table 5.7: Results from M1 by using both micro and meso

5.5.2 M2: Change Vectorisation

The second modification contains two changes to the implementation. The first changes the words used during vectorisation of the user's micro and meso vectors. The second is how terms are "counted" when calculating their vector weights. As seen in the Fernandez et al. [2018], the vectors are created using all the unique terms found in the entire dataset, referred to as W_p. The first change is to replace W_p with those found in the dictionary, L. In theory, the words from the radical dictionary are important and could reduce misclassification by removing "noise" from the nonessential words. The radical dictionary's terms created in Experiment 4 are retrieved and sorted based on their similar score. After sorting all 2764 terms, the first 305 terms are used for the experiment to replace the W_p during the vectorisation of the micro and the meso vector.

The second changes the frequency for calculating each term's vector representation. The change replaced the method with a modified one-hot-encoding approach. The frequency value is based on the number of appearances in all the posts, meaning it counts how many posts contain the term. This means that the value will still be calculated as one when the word is used multiple times in a post. The idea is instead to represent the term's existence in the post, not the frequency. This change reduces the possibility of users repeating a radical term multiple times in one post and getting a high weight for the term.

	Precision	Recall	F1	Precision	Recall	F1	AverageF1
DecisionTreeClassifier	0.654	0.624	0.592	0.661	0.630	0.604	0.598
GaussianNB	0.670	0.657	0.641	0.705	0.689	0.673	**0.657**
LogisticRegression	0.661	0.658	0.642	0.668	0.665	0.647	0.645

Table 5.8: Results from M2

As in Experiment 1 and the implementation in Fernandez et al. [2018], each user's micro and meso values were used individually to train and test the classifiers. The result in Table 5.8 (from left to right) presents the name of each classifier, together with the precision, recall, and F1 score of only using the micro vector. The next precision, recall, and F1 score are the results only using the meso value. Further, similar to testing in Experiment 1, the results of training the classier with both micro and meso values together are presented in Table 5.9.

	Precision	Recall	F1
DecisionTreeClassifier	0.709	0.677	0.659
GaussianNB	0.703	0.693	0.681
LogisticRegression	0.708	0.706	0.690

Table 5.9: Results from M2 using both micro and meso

5.5.3 M3: Adding Two Metrics

The third modification, M3, is a continuation of the second modification. This is based on the improved results in M2, seemly reducing noise during classification. The modifications in M3 include the profanity ratio and average length of posts' for each user's original and shared posts. As explained in the experiments' plan section, four new values will be created. Two will be for the original posts and be with the micro value, and the other two will be with the shared posts with the meso value. The profanity ratio will correspond to each set of posts, contributing to the classifiers with meta-data of the posts. For example, when calculating and training only the micro value, the classifier will be presented with the user's profanity ratio and the average length in the original posts. The same will be calculated using the meso value, but only for the shared posts. The reason is that the micro value shows the radicalisation on an individual level, while the meso shows it at group level. The individual level of radicalisation is measured using the posts created by the user, called the original posts. The group level is the posts shared by the user and represents the radicalisation on a group level.

The code was executed four times with combinations of using the micro value, meso value, the average length of original posts, the average length of shared posts, the profanity ratio of original posts, and the profanity ratio of shared posts. The tests are performed in the same manner as in Experiment 1, and Fernandez et al. [2018]. The classifiers are trained on the value in "micro" and then "meso". The test is also performed in a 10-fold. An additional test of each combination will use all the values, meaning not separating between the values related to the original posts (micro) and the values related to the shared posts

(meso). The test using all the values will be compared to the results from Experiment 1, shown in Table 5.2. The table will also be constructed in the same manner as in Experiment 1, where from left to right, the first column represents the classifiers used with the following precision, recall, and F1 value when using the "micro" values. When performing the tests with the new metrics, the division between micro and meso was maintained. During the use of the micro value, the average length of posts and the profanity ratio were calculated based on the original posts and used. When using the meso value, the average length of posts and the profanity ratio were calculated based on the shared posts and used. The last column, AvgF1, will average the micro F1 and meso F1 scores. The tables with this structure are Table 5.17, Table 5.15, Table 5.15 and Table 5.14.

The first combination uses "all" the new information when training the classifiers. The classifier is first presented with the user's micro value, profanity ratio, and the average length of the original posts before evaluating the test data. Furthermore, when trained in the meso value, the classifier is presented with the user's meso value together with the shared post's profanity ratio and average length value before evaluation. The results are presented in Table 5.10. The last and additional test uses all the data, meaning the user's micro value, the average length of original posts, profanity ratio in the original posts, user's meso value, average length in shared posts, and profanity ratio in shared posts. Combining all the features is used to train and test the classifiers to improve the results. The result is presented in Table 5.14.

The second combination is equivalent to the first but excludes the average length value on the posts. The profanity metrics are added during training and test both the micro and the meso values. The results are presented in Table 5.11. The additional test combines the micro, meso, and profanity ratio values in both original and shared posts and excludes the average length values. The results can be seen in Table 5.15.

The third combination is the same as the prior combination, but only that the profanity ratio is changed with the average length value. The micro and meso are trained separately but include their corresponding average length values, meaning the average length of the original post and shared posts. The results are in Table 5.16. The additional test is also performed where all the classifiers are trained with the user's micro, meso, average length on the original post, and average length of the shared post. The results from the additional test are presented in Table 5.16.

The fourth and last combination is performed as a baseline on the newly included

metrics. Like the other combinations, the classifiers are trained separately to see the performance of the classifiers to detect radicalisation at individual and group levels. The micro and the meso values are removed from the training. The classifiers are trained only on the profanity ratio and the average length of the users but are still divided between the values of the original or share posts. The results from only using the profanity ratio and average length of the original posts, then the profanity ratio and average length of the shared posts, are presented in Table 5.13. The additional test, meaning training and testing with only profanity ratios and average length values of original and shared posts, are presented in Table 5.17.

	P	R	F1	P	R	F1	AvgF1
DecisionTreeClassifier	0.831	0.837	0.829	0.897	0.887	0.884	0.**857**
GaussianNB	0.856	0.837	0.826	0.853	0.782	0.766	0.796
LogisticRegression	0.837	0.819	0.803	0.871	0.817	0.808	0.806

Table 5.10: Micro with profanities and average length of posts, and Meso with profanities and average length of posts

	P	R	F1	P	R	F1	AvgF1
DecisionTreeClassifier	0.833	0.833	0.814	0.842	0.847	0.834	0.824
GaussianNB	0.853	0.841	0.822	0.865	0.858	0.845	**0.833**
LogisticRegression	0.854	0.817	0.791	0.803	0.803	0.781	0.786

Table 5.11: Micro with profanities, and Meso with profanities

	P	R	F1	P	R	F1	AvgF1
DecisionTreeClassifier	0.584	0.58	0.574	0.828	0.824	0.812	**0.693**
GaussianNB	0.626	0.611	0.594	0.834	0.751	0.725	0.659
LogisticRegression	0.575	0.57	0.558	0.851	0.789	0.77	0.664

Table 5.12: Micro with average length of posts, and Meso with average length of posts

ML	P	R	F1	P	R	F1	AvgF1
DecisionTreeClassifier	0.801	0.805	0.8	0.883	0.872	0.873	**0.837**
GaussianNB	0.844	0.809	0.804	0.853	0.782	0.763	0.784
LogisticRegression	0.808	0.792	0.773	0.869	0.815	0.799	0.786

Table 5.13: Only profanities and average length of posts

	P	R	F1
DecisionTreeClassifier	0.948	0.945	**0.943**
GaussianNB	0.879	0.838	0.828
LogisticRegression	0.885	0.851	0.842

Table 5.14: Using both Micro and Meso with profanities and average length of posts

	P	R	F1
DecisionTreeClassifier	0.927	0.919	**0.916**
GaussianNB	0.915	0.918	0.912
LogisticRegression	0.906	0.886	0.879

Table 5.15: Using both Micro and Meso with profanities

	P	R	F1
DecisionTreeClassifier	0.803	0.819	**0.802**
GaussianNB	0.835	0.777	0.744
LogisticRegression	0.782	0.758	0.728

Table 5.16: Using both Micro and Meso with average length of posts

	P	R	F1
DecisionTreeClassifier	0.922	0.921	**0.918**
GaussianNB	0.878	0.832	0.822
LogisticRegression	0.877	0.844	0.834

Table 5.17: Using only profanities and average length of posts

Chapter 6

Discussion

This chapter discusses the results from the experiments in Chapter 5, in relation to the research questions presented in Section 1.4. In order to properly answer each research question, this chapter will be divided into sections aiming to give a response to each proposed question.

6.1 Discussion

Research question 1 will use the results of the implementation in Experiment 1. Research question 2 will use the results and analysis of the datasets in Experiments 2 and 3. Research question 3 will use the results from Experiment 4. Research question 4 will use the results from implementing the modified methods in Experiment 5.

6.1.1 Research Question 1

RQ1: *Is there a method for detecting far-right users within another domain of radicalisation detection?*

Deciding on method

The decision to use the method in Fernandez et al. [2018] was based on the findings in Chapter 3. While searching for the current state-of-the-art in detecting radicalisation, it was discovered that the research mainly focused on analysing text, not users. The method presented in Fernandez et al. [2018] (see Section 3.5)

uses the entire profile of the users on social media, using the human profile of users. This focus on user profiles is contrary to the other methods from research. Research seemed to view the task of detecting radicalisation as a binary problem of textual classification rather than a human-centered problem. As argued in Borum [2011], the mission is not on what people believe but rather how they end up with their beliefs. A framework to predict and detect radicalisation, in my view, needs to be founded in theory on the psychology or sociology of humans and not only a "computational" approach. Within the search for related work, a method proposed by Fernandez et al. [2018] stood out. The approach used theory from another science domain, social science, to create a possible solution to the problem from a new viewpoint. The task in the paper focused on the detection and prediction of Islamic radicalisation on Twitter. Based on a theory from social science named "Roots of Radicalization", the theory was "translated" to a computational approach. With the method, the users in "How Isis uses Twitter" (see Section 3.4.2) got their micro and meso influence scores calculated and used on the three classifiers. The results showed an outstanding average F1-score of 0.901 [Fernandez et al., 2018]. The theory and computation approach seem to perform very well.

Another candidate to be the primary method of the thesis was the approaches found in Lara-Cabrera et al. [2019]. With elements from a field combining psychology and linguistics, called psycho-linguistics, the approach suggested using the features of introversion and frustration in a user's text. Frustration was calculated by finding the percentage of profanities in the text, while introversion with the average length of sentences. The results were exciting and had the potential for further investigation. The problem is that the features found to be suitable for detecting is not based on fully supported by theory on radicalisation, but rather empirical evidence. That means that frustration is an excellent feature to include when detecting radicalisation, but does that mean that all frustrated people are on their way to becoming radicals? Features should be created based on a theory, not the other way around, where features create a theory or hypothesis. Using only features based on empirical evidence can increase the chance of returning false negatives. Users showing dissatisfaction online could be labelled radical by accident..

The promising results by Fernandez et al. [2018] are not only good for detecting Islamic radicalisation but indirectly support the social science theory as well. The theory is about how the process of radicalisation is a process at the micro, meso, and macro levels. It is is presented as independent of the type of radicalisation, suggesting a universal radicalisation process independent of ideology and ideas. This premise can be used to assume that similar results could be reached

by implementing it in other fields of radicalisation, such as the far-right an alt-right radicalisation investigated in this thesis. Therefore, the primary method of this thesis closely follows the methods presented by Fernandez et al. [2018] as the primary method for the thesis.

Implementation and results

The implementation in Experiment 1 was implemented as close to the description in Fernandez et al. [2018]. To implement Experiment 1, a radical dictionary and datasets as created. The dictionary would contain radical terms related to the type of radicalisation: a set of words essential for the targeted radical users. In the case of Fernandez et al. [2018], the terms were related to Islamic radicalisation. The dictionary in Fernandez et al. [2018] was constructed by combining dictionaries created by organisations who focuses on radicalisation. The task is fulfilled by Research question 2, where a method of generating terms by using manifestos was performed. The second task to fulfill was the creation of a dataset. Research question 3 with Experiments 2 and 3 created the two needed datasets, one containing radical users and one containing non-radial users. The consequence of creating both a radical dictionary and datasets was that the execution of the experiment was not in chronological order, making Experiment 1 be performed second to last. The two datasets were combined to create a dataset containing 112 radical and 112 non-radical users. The radical dictionary was sorted where the first 305 was selected.

As seen in Table 5.1, the results were achieved in this experiment were much lower than the results from Fernandez et al. [2018]. The table shows the various classifiers used for testing and training, where the first Precision, Recall, and F1 represent the results by only using the micro value, while the second represents the meso value. The best achieved result, based on the average F1-score of micro and meso, was 0.569. By analysing the image in Figure 5.1, there is no clear distinction between the radical (red) and the regular (blue) user. An ideal distribution would center the regular users around the lower left part, meaning they score low on both the micro and meso values and the radical in the far top-right corner. The results suggest the method is unsuitable for detecting far-right users, but there are two other likely reasons for the poor results; the terms and the dataset. The quality of terms is discussed in Section 6.1.3, and the quality of the dataset in Section 6.1.3.

A seemingly small "cluster" of more radical users can be seen in the top-right corner in Figure 4.2.1. The clustering seems to indicate the implementation still has some validity since almost only radical users are placed in the top-right corner. This suggests they score high on both micro and meso radicalisation scores. This could be duo to the difference in post length between a radical and a regular

user dilutes the calculated weighing of the user's micro and meso. The average length of a post by radical users was 374 characters long, contrary to a regular user, where the average was 99 characters. Length of characters likely means more words in each post, giving the radical user a higher number of "words" when calculating their vectors. The result is that radical users have a less sparse vector, and therefore, a higher chance to be more similar to the $\vec{L}$ vector. The differences in length of characters in each post is supported by the findings in Lara-Cabrera et al. [2019], where the initial idea of higher introversion in "radical" people manifested itself by having shorter posts. In the contrary, the findings of this study showed that radical users usually wrote longer texts than regular users [Lara-Cabrera et al., 2019].

A potential bias in the dataset may have contributed to the length difference in the post. The premise when constructing the radical dataset was to find people banned from Twitter, based on the comment presented in Experiment 2. To become "banned", one has to interact with the platform to make it possible to identify the posts as radical, making them somewhat interactive users. This means that the radical users banned are more likely to interact with a social media platform resulting in more content on their profiles and longer posts, contrary to the regular users. This means that the dataset of radical users does not only contain people with radical ideas, but also highly active content-contributing users to the platform. This can explain the averaged 275 character difference between regular users and radical users. As mentioned in Section 2.1.5, the phenomenon supporting this hypothesis is the 1% rule, or 90-9-1 principle, where in internet communities, 90% of users will not contribute, while the 9% occasionally contribute, while the last 1% creates the majority of content. In the study, Trevor [2014] the conclusion shows this rule of thumb is consistent in four different sites and even suggests that the 1% is far less than 1%. The radical users, because of this bias, can be retrieved from the 9% or the 1% users.

The method selected for the task seems to be a good candidate. Based on the results from Fernandez et al. [2018], the method is good at detecting radicalisation. But as Experiment 1 shows, the implementation based on the description in Fernandez et al. [2018] showed poor results. Some improvement can be seen in Experiment 5, which is going to be discussed in Section 6.1.4, but overall does not show sufficiently satisfactory results. The final thought on the usage of the model in the detection of far-right/alt-right will be discussed after investigating whether there are areas to improve. The base for "improvement" will be based on the results achieved in Experiment 1 and compared to the results in Fernandez et al. [2018].

6.1.2 Research Question 2

RQ2: *Does a dataset suitable for detecting far-right users exist? If not, how can it be created?*

A dataset mentioned the most times throughout radicalisation research was the dataset "How ISIS Uses Twitter". The dataset is mentioned in the studies Fernandez et al. [2018], Fernandez and Alani [2018], Lara-Cabrera et al. [2019], Nouh et al. [2019], and Ul Rehman et al. [2020], where it is analysed, or used to train machine learning models. Created by the digital agency Fifth Tribe, the dataset contains 17'000 tweets by more than 100 users supporting the terror organization ISIS. Besides the tweet's content, the dataset also contains metadata such as name, username, description, localisation, number of followers and posts, timestamp, and date. The data is on Islamic radicalisation, more specifically, jihadism. Since the dataset was already extensive researched, and was another sort of radicalisation, it was not used in this thesis As explained in the introduction in Section 1.1, the desire was to explore both radicalisation of far-right/alt-right and a new source of data, Gab. The idea of using Gab as a source of data is not novel, and many curated datasets are publicly available. A corpus named "Gab Hate Corpus" consisting of 27,665 posts from "Gab.com" was published in 2022 [Kennedy et al., 2022]. The posts in the corpus were labeled with a proposed heuristic labeling system to better define the different types of hateful language. The system is based on research across computer science, psychology, and social studies and was annotated by at least three trained annotators. However, contrary to the dataset used in this thesis, the dataset contained information on posts rather than users. The posts are shown one by one with the corresponding labels given by the authors. By not having the entire user's feed containing all their posts, using the method from Fernandez et al. [2018] can not be used to identify radical individuals.

Additional datasets and corpus are available but do not satisfy the two criteria; it should be focused on alt-right radicalisation and the data should be of the enitre user profile. Therefore, the task was to create a dataset containing both radical users and non-radical users. The more straightforward dataset was to create a "neutral" dataset in Experiment 3 where the data consisted of regular users on social media. As seen in the study Fernandez et al. [2018], non-radical users were extracted by randomly selecting users on Twitter. There are multiple rules on Twitter, such as not promoting violence against a person, abusive language, or unwanted sexual advances. The consequences of breaking the rules, at

worst, result in suspension[1]. Based on their moderation and popularity, the assumption is that regular people can be found on the platform, contrary to other, less moderated platforms, such as Gab. The Twitter API was used to extract people posting and was used to construct the non-radical dataset in Experiment 1.

Another dataset was needed, the one containing the radical users. The source of data was selected to be Gab. The platform Gab, as explained in Section 2.1.3, is a platform where free speech is seen as an essential value, but as a consequence of no moderation, it has become a breeding ground of hateful users. Reports show that users with neo-nazi and white-supremacist values comprise a considerable part of the users [Kennedy et al., 2022]. The belief in white-supremacist is a core value in the ideology of the alt-right [Hawley, 2018, p.11], and it is therefore safe to assume alt-right radicals can be found on the site. The dataset created in Experiment 2 was created and consisted of users on Gab who had reported in a comment section that they had been suspended from Twitter or Facebook. With the inspiration of Fernandez et al. [2018], additional confirmation by validating if the users are "suspended" was performed by using the Twitter API. Approximately three hundred users showed to be suspended. As mentioned earlier, suspension can result from breaking the rules on Twitter[2], like harassment, foul language, or promoting violence. The conclusion was that all users retrieved from Gab, where they had self-reported being banned, and the statement also got validated with the Twitter API, got labeled as radicals.

As shown in the first chapter, in Section 1.3, there exists virtually no golden standard definition of what far-right, alt-right, and extremist means. The same problem can be said when evaluating the content of a dataset on how "far-right" or "radical" it is. It was, therefore, difficult to decide how to measure radicalisation in a dataset. The approach to evaluating the result from experiments 2 and 3 is to analyze the usage and pattern seen in the data, specifically, the usage of hashtags and metadata. One key difference between the two datasets was the usage of hashtags. In the non-radical dataset, the number of hashtags found was 13742. Divided by 213 users, the average number of the hashtag per user was 64.51. In the radical dataset, the number of hashtags found was 210482, with an average of 723 hashtags per user. Similar conclusions were drawn in the findings in Ferrara et al. [2016] and Nouh et al. [2019]. In the study by Ferrara et al. [2016], it was concluded that, the ratio of retweets to tweets, the average number of hashtags adopted, the sheer number of tweets, and the average number

[1] `https://help.twitter.com/en/rules-and-policies/twitter-rules` [Accessed on 07.07.2022]

[2] `https://help.twitter.com/en/rules-and-policies/twitter-rules` [Accessed on 07.07.2022]

of retweets generated by each user, rank exceptionally high in predictive power when classifying ISIS radicals. In Nouh et al. [2019], in the tenth position of the list of most important features in the study, the number of hashtags was concluded to be a good feature. The results of the Ferrara et al. [2016] and Nouh et al. [2019] show similar findings as those found in this thesis's radical dataset created in Experiment 2. These shared findings can indicate the same conclusion that the users in the dataset are radical users. There are also differences between the regular and radical users in the usage of hashtags. In Table 5.3, the most common hashtag of radical users can be seen. The table contains the ten most frequent hashtags, while the rest of the hashtags can be seen in the word clouds in Figure 5.2 and in Figure 5.2. The first word cloud represents the usage of hashtags found in the original posts made by radicals.

Furthermore, by analysing the patterns of hashtags and comparing radical and regular users, the hashtags are clearly different. The radical users seem to use hashtags more related to politics, contrary to regular users, who seem to be related to pop-culture, such as the Korean band BTS[Skagseth, 2021]. The important hashtag in radicals users' original posts users were #democrats, #trump2020, #joebiden, #cdnpoli, #maga, #trudeau, #covid, #covid19, #trump and #votetrump. Both Joe Biden and Trump were candidates for the election in 2020. Trump represented the republicans while Joe Biden represented the Democrats. A famous slogan by Trump when he ran in 2018 was Make America Great Again, where the arbitration is MAGA. Both #covid and #covid19 refer to the novel COVID-19 pandemic. The frequent use can be supported by the suggestion in the report Gjørv [2012]. They theorised that the far-right would catch popularity as a result of the public's dissatisfaction with the handling of the virus. The hatred toward the government was thus amplified. By the regular user's original posts, the hashtags #iheartawards, #butter, #bestmusicvideo, #bestfanarmy, #btsarmy, #bts_butter, #bts, #footiestories, #gayc and #bitcoin. Almost all ten hashtags are related to BTS, the K-pop band[Skagseth, 2021, p.12], where there are indications of high activity by the BTS's Army, the collective name of the fanbase[Skagseth, 2021, p.20], around an award. There are references to their music, the award, and the band itself by the users.

There is room for concluding the quality of the dataset. Based on the different sources where one is labeled as a radical site, Gab, while the other is one of the most popular sites on the web, Twitter, there are high chances of users sharing the views. The number of hashtags supports the idea that the users of Gab are radical-based on the number of hashtags. Furthermore, the analysis of hashtag users suggests that the radical users are more focused on politics and the pandemic, while the regular users were talking about a Korean pop band.

The final results can be suggested will manifest them-self in Experiment 1 and Experiment 5, where a model is implemented to train and test classifiers. Poor results can indicate the unsatisfactory quality of the datasets. The final thought will be discussed in Section 6.1.4.

6.1.3 Research Question 3

RQ3: *What are the advantages and disadvantages of using an unsupervised method to find radical terms?*

Human intervention is common when a dictionary of terms is required for term-based methods. As in Fernandez et al. [2018], the approach of creating the set of radical Islamic glossaries for the experiment is created by combining four different previously created dictionaries. Three originated from domain experts in institutions/departments such as Saffron and ICT, who have been working on radicalisation. The last originated from the study in Saif et al. [2016]. A less human-dependent method can be observed in the study by Ul Rehman et al. [2020], where terms were extracted from 17350 Tweets of radical and non-radical users by weighing the terms with TF-IDF and extracting the highest weighted terms. Additionally, the approach of the Saffron EU project to create a list of radical Islamic terms, the most common terms in ISIS propaganda magazines, was used to create the dictionary. In other words, the highest frequency terms were utilised to develop the dictionary by Saffron EU in Fernandez et al. [2018].

The approach proposed and performed in Experiment 4 is an unsupervised method to extract radical terms by utilising far-right/alt-right manifestos. The hypothetical advantages are removing human bias/errors and the potential for detecting "new" terms. Biases in datasets and human annotation is a field of research currently explored. The consequence of biases in machine learning is lower fairness and inferior performance[Al Kuwatly et al., 2020, p.1]. As explored in Al Kuwatly et al. [2020], differences in demographics characteristics between humans can have effects when annotating data. In Al Kuwatly et al. [2020] an experiment where diverse groups annotated the corpus from Wikipedia's Detox project, there were differences in labeling correlating to demographical characteristics. Features such as if the person is a native English speaker showed identified personal attack in comments better than non-native English speakers. The suggestion is that native speaker better detect personal attacks in comments. Another correlation was that the differences in age and education lever affected the "results" of the labeling. Humans performing the labeling task showed variation in annotation, which can be an unexpected challenge when creating systems for the detection of radicalisation. By utilising machine learning models, more specifically unsuper-

vised models, a system can work without human differences, potentially creating better quality terms. Noteworthy, I will not claim that no biases exist in pre-trained Transformers models. The method used in Experiment 4 is KeyBERT, which utilises the transformer model BERT (see Section 2.5.3). As discovered in the study Bhardwaj et al. [2021], BERT does contain a gender bias in the vectorisation. That can be one of many biases found in the Transformers models worth considering.

The further advantage of using the approach with the manifestos is the change for the system to keep up to date with alt-right online. Alt-right and far-right users are known for using "secret" language, as triple parentheses around a name to mark the person for having Jewish heritage or number as 88. The numbers represent the position of letters in the alphabet, where "88" means "HH" referring to "Heil Hitler" (see Section 2.1.3).

As seen in Table 5.5, the terms extracted from the manifesto are understandable and, by inspection, related to ideas of the far-right and alt-right. The different explanation in the following manifesto is based on the content of the manifesto and was analysed by the author in this thesis. The first manifestos extracted terms relating to different ideologies. Marxist and Marxists refer to the ideology Marxism, which is on the left side of the political spectrum (see Figure 1.1). As explained in Section 1.3.3 and shown in Figure 1.1, the right and left are placed in two different places in the spectrum where they can be understood as almost contradictory political ideologies. So, naturally, rivalry and hate for the ideology opposed to the far-right are despised. The second manifesto contains terms such as society, dealing, suffer, inferior, blackness, manifesto. The manifesto explains in detail the theory about "The Great replacement" and talks about how the white race is getting replaced by "inferior people" who are, according to the manifesto, behind much of the crimes as "dealing". The manifesto mainly focuses on people of African dependence, which is the reason behind the term "blackness". The third manifesto, Manifesto 3, shares a lot of the same view as Manifesto 2. However, the difference is that "his inferior people" in The Great Replacement are the people of Hispanic dependence in Texas. Manifesto 4 has a different structure than the rest. The manifesto is structured as a book containing the recipes for the used weapons for the terror event and a form of "gamification". The last pages contain a list of points acquired behind achievements, similar to games. Due to the manifest's descriptive nature, many terms related to guns and weapons were detected by KeyBert.

The terms found in Experiment 4 seem to represent the different manifesto well. Future work can explore if the terms can be selected adequately by using terms

crossing with other manifestos or analyse if terms get better by adding more manifestos to the system. In conclusion, both advantages and disadvantages exist of using an unsupervised method to select terms. The trade-off of less human interaction can hence mean hidden biases or poorer quality in terms. The radical dictionary, created in Experiment 4, was tested when used in Experiment 1 and 2, and is further discussed further in Sections 6.1.1 and 6.1.4.

6.1.4 Research Question 4

RQ4: *How can the existing approach for detecting radicalisation be improved?*

The implementation performance in Experiment 1 following the steps in Fernandez et al. [2018] showed poor results. With the desire to improve the results, small suggestive changes were made to the model. The improvement will be compared to the achieved performance in Experiment 1, as seen in Table 5.1 and Table 5.2, and the results from Fernandez et al. [2018] will also be discussed, as shown in Table 4.1. To summarise the different tests, in Fernandez et al. [2018] the system was tested with three classifiers: a decision tree, a naive Bayes, and a logistic regression model. The system was trained and tested separately with the micro and the meso values. The metric calculating performance was with precision, recall, and F1, and the test was performed in a 10-fold. Additionally to the tests performed in the study, an additional proposed test was added in Experiment 1, shown in Table 5.2. The experiment uses both values during training and testing, meaning the users' micro and meso values. The idea is that by providing more user information, the better the system can learn to classify the users as radical or not. To summarise the three different aspects and suggestions to improve in Experiment 5, the first is about adding more radical terms to the system, the second on reducing noise, and the last is about adding two metrics.

The first modification, M1, changes the number of radical terms used in the system from 305 to 2764. The hypothesis is that by including more of the terms found in Experiment 4 and by KeyBERT, the system can better calculate the radicalisation of each user at both micro and meso levels. By increasing the number of terms, the performance could improve by considering more of the terms found by KeyBERT and potentially create a more apparent distinction between the radical and non-radical users. As seen in Figure 5.1, the red(radical users) and blue(ordinary users) users are all scrambled up, making it challenging for a classifier to detect and classifier users. A semi-clustering can be seen in the top right corner, where ideally, all radical users would be found. To be placed in the top right corner, the users must score high on similarity with the selected terms from the radical dictionary. This means the 305 terms originally, in both

the original and shared posts. By changing the number of terms, the change increases focus on the more hidden important terms. However, the commentary can result in more noise-reducing in the system's overall quality.

The second modification, M2, experiments with reducing the terms used when vectorising. The implementation from Experiment 1 and Fernandez et al. [2018] uses all the unique terms found in the entire dataset, called W_p. A challenge is increasing unique words when implementing more data and users to the system and creating more sparse matrices when calculating the micro and meso values. The idea is, based on the theory behind Experiment 4, the terms selected by KeyBERT are closer to the essence of radicalisation. By focusing on the terms found in the implementation, the system should potentially reduce noise and only focus on the important terms. The implementation uses the terms retrieved by KeyBERT in Experiment 4 and sorts them based on their similarity. The first 305 terms are selected and used to calculate the micro and meso values.

The last modification, M3, implement two new metrics based on empirical evidence from Lara-Cabrera et al. [2019]. In the study performed in Lara-Cabrera et al. [2019], different indicators of users are evaluated if they are suitable for detection radicalisation. In the study, the test data "How-isis-used-twitter" is used and evaluates five different keyword-based approaches to detect text if a user is radicalised. Of the five indicators analysed in the study, two essential features stood out: frustration and introversion. The belief is that a user that expresses frustration conveys dissatisfaction toward the society in the form of either capitalised text or containing more swearing than ordinary users. The second is introverted. The idea supported by other research is that introverted people manifest it by writing text shorter than regular people [Lara-Cabrera et al., 2019, p.972-973]. The results in that study supported the hypothesis about frustration but showed the contrary to inversion. The users labeled as radicals, in the case of the study ISIS-supportive, in reality, tend to write longer than the standard user. It can be argued that the feature can still be vital since it can help distinguish people with longer text and potentially correctly identify radical users with the other features. A similar finding is shown in Grover and Mark [2019]; users from and alt-right subreddit show more negative scores than regular users. Based on comments from different subreddit over a period spanning six months, the text analysis showed that racial slurs, antisemitism, and politically related terms were the highest-scoring terms based on their TF-IDF score. Additionally, user's profile scored higher in more negative emotions, meaning the user's text was dominantly negative emotions. It suggests that profanity is even more relevant concerning far-right and alt-right detection. As presented in Section 1.3.1 and Section 1.3.2, the definitions of radicalisation and extremism can be used

to explain an ideology or belief contrary to the core values in the society or a need for a radical change in society. This can imply the increased frustration of extremists and radicals because the current status of society differs from their ideal worldview.

The first modification worsens the results of the system. The highest acquired results were 0.520 in average F1 score, as shown in Table 5.6. When using both micro and meso together during training and test, the results jumped to 0.556. As shown in Figure 5.6 and Figure 5.7, the regular users are more scrambled when including all the terms. The plots of the regular and radical users get even blurrier. The worsening of the results can be a consequence of two different reasons. Either the terms are not as useful and should be inspected by humans before using, or that it exists an optimal percentage of terms from each manifesto where including more terms will result in worse performance. Future research should research the correlation between the number of terms and the system's performance using terms extracted by an unsupervised method such as KeyBERT. The poor results from M1 consequently delayed the usage of all the terms for the model presented in this thesis. Besides returning bad results, the system also used more time for the calculation of micro and meso. Consequently, the used terms in M2 and M3 were set to only consist of the standard 305 terms.

In the second modification, M2, the system was set only to use the radical terms when vectorising, meaning calculating the micro and meso values. It used the 305 terms and returned better results than Experiment 1. As seen in Table 5.8, the highest average F1 score was 0.657, increasing the results by 0.088. The results are still below the original implementation in Fernandez et al. [2018], but an increment related to Experiment 1. All three classifiers increased overall with the results, potentially showing that reducing the focus on all unique terms and only on the radical terms can reduce the noise in the system. The results from using both the micro and meso values, shown in Table 5.9, also showed an increment with the highest F1 value of 0.690. The system seems to improve the implementation by only focusing on radical terms. The results can enforce the idea that the system improves by narrowing down the number of words vectorised. Due to the improvement, the systems vectorisation method was further used in the modification of M3.

The third modification, M3, experiments with including a profanity ratio and average length value to both the user's original and shared posts. The belief is that the frustration and longer posts will help the system detect radicals. Additionally, four combinations were performed to compare the results, where the last one excluded the micro and meso values. The idea behind it, is that one can investigate which of the two new metrics are doing most of the work. The combination using all information, meaning the new metrics with their micro and meso scores, returns the best averaged results found in all experiments, with an 0.857 average F1 score. As performed in the other experiments, the micro and meso values together were used to train the classifier and test its performance by

combining all the information. In this case, the new four metrics were included in addition to the micro and meso. The result are presented in Table 5.14, and the highest F1 score with 0.943. The results surpass the 0.901 in Fernandez et al. [2018]. The results support that the frustration and average length of posts are good indicators, as mentioned in Lara-Cabrera et al. [2019]. Furthermore, combining the two theories from Lara-Cabrera et al. [2019] and Fernandez et al. [2018] returns even more promising results. Based on the promising results, the plan was to explore which metric improved the system the most; therefore, three additional tests were performed. The following test experimented with the micro and meso values, only the swearing metric, only the average length, and, lastly, both metrics.

The first combination only includes the swearing ratio together with the micro and meso values, where the highest achieved results were 0.833 on average F1 score, as shown in Table 5.11. The results are higher than the one accomplished in Experiment 1 but still lower than the 0.901 from Fernandez et al. [2018]. Combining the micro, meso, and the two swearing ratios, the system accomplished an F1 score of 0.916, as presented in Table 5.15. The next combination removed the swearing metric and only included the average length with the user's micro and meso scores. The highest accomplished average F1 score was 0.693, as shown in Table 5.12. This change is a worsening of the results accomplished by the two prior combinations but still an improvement of the system. The average seems to imply that the feature is less critical than the swearing ratio but still contains a form of importance, as seen in the first combination. Neither the swearing ratio of the average length value alone achieved as promising results as including all metrics, as preset in the first combination (see Table 5.10 and Table 5.14). The average post length and swearing ratio seem to be good together when classifying the radicals. By running the test again and including the micro and meso with their average lengths of the post, the best F1 score achieved was 0.802, as shown in Table 5.16.

The last combination removed the micro and meso scores calculated. The purpose was to explore if the swearing and average length are by themselves better metrics than the micro and meso values from Fernandez et al. [2018]. The highest achieved classification when only using the swearing and average length values was 0.837, as presented in Table 5.13. When only using the four metrics, meaning two swearing and two average length values, the results showed, at best, an F1 score of 0.918. The result is higher than a few other combinations and even the results in Experiment 1. The features alone are good to classify, but as shown in Table 5.14, combining all the metrics with the micro and meso value, the performance is still below the 0.942 F1 scores.

The overall results from all the experiments imply and represent the possibility of employing methods from other radicalisation research in new fields. These results shown in the experiments indicate that implementing borrowed methods can be crucial for both developing and creating systems for far-right and alt-right detection.

Chapter 7

Conclusion and Future work

The goal of the Master's thesis was to detect far-right and alt-right radicalisation on social media. Four research questions were formulated to cover the research goal. The first research question, RQ1, focuses on finding a method with the potential to be tailored and implemented for detecting far-right radicalisation. The selected method is referred to as the primary method or just method. The second research question, RQ2, focused on constructing datasets for detecting far-right radicalisation. Two datasets were needed: one with radical user and another with regular users. Radical users are far-right users, and regular users are ordinary users on social media platforms. The datasets are referred to as the radical dataset and the non-radical dataset. The third research question, RQ3, focused on creating a method for creating radical terms. Radical terms are terms associated with far-right users online. A set of all the terms is referred to as the radical dictionary. The fourth research question, RQ4, focused on using the primary method and improving it. Three suggestions were made: the first modification adds more radical terms to the method, while the second changes the method of vectorisation during the calculation of micro and meso. Micro and meso are a reference to the different levels of radicalisation in Fernandez et al. [2018] (see Section 3.5). The third modification adds two new metrics, the profanity ratio and average length of the posts.

Five experiments were designed to answer the different research questions. RQ1 has Experiment 1, RQ2 has Experiment 2 and Experiment 3, RQ3 has Experiment 4, and RQ4 has Experiment 5. Experiment 1 is about implementing the methods selected in RQ1 and testing the method. The tailored implementation uses the radical dataset, the regular dataset, and the radical dictionary from the later experiments. Experiment 2 focuses on creating a dataset containing far-

right users. Users were extracted from Gab, a far-right site, to be the users in the radical dataset. Experiment 3 is similar but creates a dataset with regular users. The uses were extracted from Twitter by randomly selecting users. The dataset with the regular users is referred to as the regular dataset. Experiment 4 investigates how to extract far-right radical terms. The terms were extracted from manifestos by extremists. The set of terms is called the radical dictionary.

The following section will go into depth and conclude the different research goal and questions of this Master's thesis.

Research Goal: Detecting political radicalisation of users on social media.

The promising results achieved by using a tailored method show opportunities for detecting far-right users on social media. By investigating the years of research on radicalisation, methods for detecting far-right users with few adjustments can be implemented. Further investigation is needed in the field using the more state-of-the-art technology.

Research Question 1: Is there a method for detecting far-right users within another domain of radicalisation detection?

For research question 1, the goal was to find an adequate model or theory that could detect far-right and alt-right radical users. Based on the discovered methods found in the Structured Literature Review in Chapter 3, one method stood out. The model explained in Fernandez et al. [2018] was selected because of its good performance and theory rooted in social science. Its performance suggested that the theory based on radicalisation as a process will also work when implemented in far-right detection. This method became the primary method used for the thesis. Two datasets and radical terms were needed to adapt the method to far-right and alt-right detection. The implementation of the model was performed after the creation of both the datasets and dictionary of radical terms. The results were worse than anticipated, with an average F1 score of 0.569. The average F1 score means the average of both the micro and the meso based on their achieved F1 score. The results were much lower than those in Fernandez et al. [2018], where they accomplished an average F1 score of 0.901. Besides the results, when users were plotted on a graph, there was a clustering of radical users in the top-right corner. The clustering indicates that the method can be used to detect far-right users because the users need to have both high micro and meso to get this placement. This means that radical users are creating and sharing content with high similarity to radical terms. It was argued that the cause could be that radical users have longer posts and, therefore, use more words.

This can be a form of bias and affect the micro and meso vector when calculating.

To conclude, implementing methods from different fields of radicalisation need tailoring. The results show poor results when used without any modifications. But as later explored in RQ4, the results can be improved with some modifications. The conclusion is that there are possibilities to use the method, but it requires tailoring.

Research Question 2: Does a dataset suitable for detecting far-right users exist? If not, how can it be created?

In research question 2, the goal was to find out whether there are datasets that could detect far-right radicalisation. There were no adequate far-right datasets based on the literature in Chapter 3 or other research with public datasets. The task changed to create two datasets containing radical and regular users. The datasets were created by extracting users on the site Gab and the regular users from Twitter. Due to Gab's reputation for containing far-right users, the dataset was assumed to be radical. In comparison, Twitter included more regular users. The challenge was to see if users were radical or not, and based on their use of hashtags; the results seemed promising. Regular users talked about music and a boyband, while the radical users talked more about politics.

The quality of the dataset were concluded to be inconclusive. This is because poor results were achieved in Experiment 1 while Experiment 5 with some modification achieved good results. In Experiment 1 the tailored implementation gave an average F1 score of 0.569, while Experiment 5 gave 0.942. The conclusion is that it can be created and that Gab is a sufficient data source for retrieving far-right users.

Research Question 3: What are the advantages and disadvantages of using an unsupervised method to find radical terms?

In research question 3, a method for creating a list of terms related to far-right and alt-right content was needed. The list, known as the radical dictionary, was created by an experimental approach suggested in Experiment 4. The idea was to use manifestos by far-right extremists to extract the best keywords. The keyword extraction will, in theory, select the best terms to describe the document. Based on the content in the manifestos, the words will also represent far-right content. Therefore, the terms will, in theory, be far-right terms. The proposed approach will reduce human biases but replace them with other biases. BERT's biases can be transferred to KeyBERT, since it uses the embedding from BERT.

In contrast, advantages can be that the system detects unknown terms. Far-right and alt-right users usually hide their language by using secret codes. Codes such as 88 ("Heil Hitler") and triple parentheses to mark people of Jewish heritage have been used by far-right users.

Based on the analysis of the terms, each selected word seemed to correlate with the content of the manifestos. The manifestos contain anti-semitism, racism, terrorist planning, and violent and extremist theories. The quality of the terms seems to be neither good nor bad. Experiment 1 showed bad results, while Experiment 5 got higher than the results in Fernandez et al. [2018]. The concluding of the quality of the terms are inconclusive.

Research Question 4: How can the existing approach for detecting radicalisation be improved?

Research question 4 investigates opportunities to improve the model. The poor result in Experiment 1 showed that the model needed modifications to better distinguish between radical and regular users. Three different modifications were suggested. The first included all the terms from the radical dictionary. The second changed the method of vectorisation, where only words from the radical dictionary were vectorised. The third modification included two new metrics, a swearing ratio, and an average length metric.
The first modification produced worse results, with an average F1 score of 0.520. Based on the graphs with and without the modification, the hypothesis is that it introduced more noise to the system. The second modification replaced the words used to vectorise the user's posts. The original model uses all unique words found in the dataset. The new modification suggests only using the terms from the radical dictionary. The results improved, resulting in an average F1 score of 0.657. The results are improved but not close to the results from Fernandez et al. [2018]. Due to the improving results, the implementation was further used in the third modification.
The third modification added two new metrics: profanities and average length of posts. The profanity ratio is the number of profanities divided by the total words. The average length is the average length of characters. Both metrics are relative to when micro or meso is used. When calculating micro, the two metrics represent the original posts by the user. With meso, the number represents the shared posts. There were four different combinations performed with the metrics. The first with profanities and average length of posts, together with micro and meso. The results achieved the highest average F1 score at 0.857. When all metrics, meaning both the micro and meso and their two metrics, the value reached an F1 score of 0.947. This was the highest achieved result. The second combination

included only the profanity rations. When used with micro and meso, the results were an average F1 of 0.833. When all metrics in this combination were used, the results reached an F1 score of 0.916. The third included only the average length metrics. The result with micro and meso was an average F1 score of 0.693. When all metrics from the combination were used, the F1 score became 0.802. The results show that the profanity ratio was the most important of the two metrics. The fourth and last combination gave a baseline for evaluating the two metrics by themself. The test was performed in the same manner, only that the micro and meso were removed, resulting in an average F1 score of 0.837. When all metrics were combined (not micro and meso), the F1 score became 0.918.

In conclusion, the best modification to the method was to include the two new metrics. The metrics were profanity and average length. The results show there is room for including more features to improve the method, which is worth further investigation.

7.1 Contributions

This thesis contributes to the field of detection of far-right radicalisation on social media. The first contribution is an alternative path to take on the challenge of detecting far-right radicals. The approach for this thesis used methods from other types of radicalisation detection but adjusted them to far-right radicalisation. The results showed promising results with some modifications, such as adding metrics and reducing the occurrence of non-relevant words. The future of radicalisation detection is to use the already well-established method found in other fields of radicalisation.

The second contribution to the field is an overview of the status quo on radicalisation detection online. The method used to extract papers (Structured literature review form Kofod-Petersen [2015]) returned multiple papers with valuable findings and methods for detecting radicalisation. The papers are not explicitly focused on far-right radicalisation. The research in this field is still too small to contain only far-right radicalisation papers. The overview contains good research papers together with promising findings.

A third contribution is an approach to creating radical datasets. The radical users were, in this case, far-right and alt-right users. They were extracted from the far-right site Gab. They also self-reported being "banned" from Twitter. All users were controlled to be banned, and excluded from the dataset if they were, in fact, banned. The final results gave 291 users. The results and analysis based on similar findings from other papers confirms that the content stems from radical

users.

The fourth contribution is the method of creating radical terms. Radical terms in this thesis are terms commonly used by far-right users. The experimental approach used the manifestos published by far-right extremists to extract important keywords. The idea is that since the manifestos contain far-right content, the terms would also represent the far-right ideas. The unsupervised method of KeyBERT was used to extract terms. The final number of terms was 2743. This method can be used without human interference and yielded good results when used in the experiments.

The fifth contribution is a tailored method based on Islamic radicalisation, which can now be used for far-right detection. The method was tailored by creating a dataset and a set of radical far-right terms. The results show the potential of improving the method by adding features supported by studies.

7.2 Limitations

In Experiment 2 and 3, assumptions were made. During the creation of the radical dataset, the users were all defined as far-right or alt-right users based on their presence on the Gab platform. This simplification of a complex task of defining what radical means, was done to reduce data collection time. No domain experts in the field were available to help evaluate the dataset. As mentioned in Section 1.3.1, radicalism is defined in various ways in the research, and there is no universally accepted definition. The simplification helped the development and implementation of the experiments but can have come at the cost of data quality. There was nothing to control if all users were radical, making the dataset's quality inconclusive.

The creation of the dataset with regular users also contained assumptions worth mentioning. The dataset's creation consisted of extracting users from the Stream endpoint in the Twitter API. The short time span using the API shows to have extracted users related to a music event. The users seem to overrepresent the Korean boyband BTS, which can indicate there was an event during the period of extraction. This may have created problems in the dataset in two ways. The first is that the users may have been created only for the event. If the event contained any award nominations through Twitter, many of the users could be "new" users only created for voting. New users could result in users with uncommonly few posts and fans only posting about the band. Secondly, the event may have concentrated the fans' presence on the platform. The concentration of the fans may have resulted in the dataset containing more BTS fans than regular

users.

A limitation that can be seen in the discussion section when evaluating the quality of Experiments 2, 3, and 4. Two datasets were created in experiment 2 and 3, which were used in experiments 1 and 5. Experiment 4 created a list of radical terms with a new approach with manifestos. Only experiment 1 and experiment 5 are evaluated based on results from classifiers. Given the multiple steps of creating the data for the implementations, deciding what is good quality is challenging. The many steps make it challenging to conclude the quality of the datasets and radical terms, which is why only the first and last experiment contain more specific conclusions.

7.3 Future Work

The following section will present suggestions for future work worth investigating. The suggestions have appeared during the coding and writing of the Master's thesis.

7.3.1 Number of Terms

The number of terms used in the experiments was 305 and originated from the study Fernandez et al. [2018]. The same number was used to justifiably compare the results by conducting the method as in Fernandez et al. [2018]. But as Experiment 4 showed, the implementation performance varies based on the number of terms. Using 305 terms gave an average F1 score of 0.569, and 2764 terms gave 0.520. Future work could explore the correlation between results and the number of terms.

7.3.2 Prediction

The original implementation in Fernandez et al. [2018] was also used for predicting Islamic radicalisation. The users' vectors were predicted/filled based on the similarity with other users. The technology for filling in the empty places in the users' vectors is from the recommendation systems. Because of lack of time, there was no focus on investigating the system's predictive potential. In theory, by adding two metrics to the method (Experiment 4), there is more information for the recommendation system. Future research should investigate using a newer method from recommendation systems.

7.3.3 Developments on Unsupervised Term-extraction

As briefly mentioned in Section 6.1.3, further investigation is needed in unsupervised methods for creating radical terms. The KeyBERT used in this Master's thesis performed its calculation based on the vectors from BERT. BERT is a Transformer model that was published and created in Devlin et al. [2018]. Since then, multiple new pre-trained Transformers have been made, such as GTP-2 and GTP-3. Thus, for future work, more comparative studies are necessary.

Bibliography

Agarwal, S. and Sureka, A. (2015). Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8956, pages 431–442. Cited By :56 Export Date: 24 November 2021.

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

Alghofaili, H. and Almishari, M. (2018). Countering terrorism incitement of twitter profiles in arabic-context. In *21st Saudi Computer Society National Computer Conference, NCC 2018*. Cited By :1 Export Date: 24 November 2021.

Ali, S., Saeed, M., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., and Stringhini, G. (2021). Understanding the effect of deplatforming on social networks. pages 187–195. cited By 4.

Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., Abushariah, M., and Alfawareh, M. (2021). Intelligent detection of hate speech in arabic social network: A machine learning approach. *Journal of Information Science*, 47(4):483–501. Cited By :5 Export Date: 24 November 2021.

Alvari, H., Sarkar, S., and Shakarian, P. (2019). Detection of violent extremists in social media. In *Proceedings - 2019 2nd International Conference on Data Intelligence and Security, ICDIS 2019*, pages 43–47. Cited By :9 Export Date: 22 November 2021.

Andersen, J. C. and Sandberg, S. (2020). *Terrorism & Political Violence*, 32(7):1506–1526. Name - Islamic State of Iraq & the Levant–ISIS; Copyright - ©2018 Taylor & Francis; Last updated - 2022-05-07.

Benigni, M. C., Joseph, K., and Carley, K. M. (2017). Online extremism and the communities that sustain it: Detecting the isis supporting community on twitter. *PLoS ONE*, 12(12). Cited By :49 Export Date: 22 November 2021.

Bhardwaj, R., Majumder, N., and Poria, S. (2021). Investigating gender bias in bert. *Cognitive computation*, 13(4):1008–1018.

Borum, R. (2011). Radicalization into violent extremism i: A review of social science theories. *Journal of Strategic Security*, 4(4):7–36. Copyright - Copyright Henley-Putnam University Press Winter 2011; Last updated - 2014-10-31.

Brzuszkiewicz, S. (2020). Jihadism and far-right extremism: Shared attributes with regard to violence spectacularisation. *European view*, 19(1):71–79.

Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E. D., Stringhini, G., and Vakali, A. (2017). Measuring #gamergate: A tale of hate, sexism, and bullying.

Chatzakou, D., Soler-Company, J., Tsikrika, T., Wanner, L., Vrochidis, S., and Kompatsiaris, I. (2020). User identity linkage in social media using linguistic and social interaction features. In *WebSci 2020 - Proceedings of the 12th ACM Conference on Web Science*, pages 295–304. Cited By :1 Export Date: 24 November 2021.

Chelvachandran, N. and Jahankhani, H. (2019). A study on keyword analytics as a precursor to machine learning to evaluate radicalisation on social media. Cited By :1 Export Date: 29 October 2021.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Deb, K., Paul, S., and Das, K. (2020). A framework for predicting and identifying radicalization and civil unrest oriented threats from whatsapp group. In *Advances in Intelligent Systems and Computing*, volume 937, pages 595–606. Cited By :2 Export Date: 22 November 2021.

Derbas, N., Dusserre, E., Padró, M., and Segond, F. (2020). Eventfully safapp: hybrid approach to event detection for social media mining. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):87–95. Cited By :4 Export Date: 22 November 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Fernandez, M. and Alani, H. (2018). Contextual semantics for radicalisation detection on twitter. In *CEUR Workshop Proceedings*, volume 2182. Cited By :1 Export Date: 22 November 2021.

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. Cited By :23 Export Date: 22 November 2021.

Ferrara, E., Wang, W. Q., Varol, O., Flammini, A., and Galstyan, A. (2016). Predicting online extremism, content adopters, and interaction reciprocity. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10047 LNCS, pages 22–39. Cited By :64 Export Date: 22 November 2021.

Fifth Tribe (2019). How isis uses twitter.

Géron, A. (2019). Hands-on machine learning with scikit-learn, keras, and tensorflow : concepts, tools, and techniques to build intelligent systems.

Giarelis, N., Kanakaris, N., and Karacapilidis, N. (2021). A comparative assessment of state-of-the-art methods for multilingual unsupervised keyphrase extraction. In *Artificial Intelligence Applications and Innovations*, IFIP Advances in Information and Communication Technology, pages 635–645. Springer International Publishing, Cham.

Gjørv, A. B. (2012). Rapport fra 22. juli-kommisjonen : oppnevnt ved kongelig resolusjon 12. august 2011 for ågjennomgå og trekke lærdom fra angrepene på regjeringskvartalet og Utøya 22. juli 2011 : avgitt til statsministeren 13. august 2012.

Gjørv, A. B. et al. (2021). Nasjonal trusselvurdering i 2021. A national risk analysis of the Norwegian kingdom by PST.

Glitsos, L. and Hall, J. (2019). The pepe the frog meme: an examination of social, political, and cultural implications through the tradition of the darwinian absurd. *Journal for Cultural Research*, 23(4):381–395.

Goodfellow, I. (2016). Deep learning.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodwin, J. (2021). Gab: Everything you need to know about the fast-growing, controversial social network. https://edition.cnn.com/2021/01/17/tech/what-is-gab-explainer/index.html.

Grootendorst, M. (2020). Keybert: Minimal keyword extraction with bert.

Grover, T. and Mark, G. (2019). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, pages 193–204. Cited By :14 Export Date: 22 November 2021.

Hall, M. (2021). Facebook - american company. *Encyclopedia Britannica*.

Hartung, M., Klinger, R., Schmidtke, F., and Vogel, L. (2017). Identifying right-wing extremism in german twitter profiles: A classification approach. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10260 LNCS, pages 320–325. Cited By :8 Export Date: 22 November 2021.

Hawley, G. (2018). Making sense of the alt-right.

Hebb, D. O. (1949). The organization of behavior : a neuropsychological theory.

IBM (2022). What is logistic regression?

Jasser, G., McSwiney, J., Pertwee, E., and Zannettou, S. (2021). 'welcome to #gabfam': Far-right virtual community on gab. *New Media & Society*, page 146144482110245.

Jenkins, B. M. (2022). Domestic violent extremists will be harder to combat than homegrown jihadists. `https://www.rand.org/blog/2021/02/domestic-violent-extremists-will-be-harder-to-combat.html`(Accessed: 27.06.22).

Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval: 60 years of the best in information research. *Journal of documentation*, 60(5):493–502.

Kemp, S. (2022). Digital 2022: Global overview report. `https://datareportal.com/reports/digital-2022-global-overview-report` (Accessed: 27.06.22).

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., Hoover, J., Azatian, A., Hussain, A., Lara, A., Cardenas, G., Omary, A., Park, C., Wang, X., Wijaya, C., Zhang, Y., Meyerowitz, B., and Dehghani, M. (2022). Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Kofod-Petersen, A. (2015). How to do a structured literature review in computer science. Technical report.

Kor-Sins, R. (2021). The alt-right digital migration: A heterogeneous engineering approach to social media platform branding. *New Media & Society*, page 146144482110388.

Kostakos, P., Nykanen, M., Martinviita, M., Pandya, A., and Oussalah, M. (2018). Meta-terrorism: Identifying linguistic patterns in public discourse after an attack. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, pages 1079–1083. Cited By :2 Export Date: 24 November 2021.

Kulkarni, A. V., Aziz, B., Shams, I., and Busse, J. W. (2009). Comparisons of Citations in Web of Science, Scopus, and Google Scholar for Articles Published in General Medical Journals. *JAMA*, 302(10):1092–1096.

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. Cited By :11 Export Date: 22 November 2021.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

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. Cited By :5 Export Date: 22 November 2021.

Mashiloane, L. and Mchunu, M. (2013). Mining for marks: A comparison of classification algorithms when predicting academic performance to identify "students at risk". In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8284 of *Lecture Notes in Computer Science*, pages 541–552. Springer International Publishing, Cham.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Minsky, M. (1988). Perceptrons : an introduction to computational geometry.

Miranda, E., Aryuni, M., Fernando, Y., and Kibtiah, T. M. (2020). A study of radicalism contents detection in twitter: Insights from support vector machine technique. In *Proceedings of 2020 International Conference on Information*

Management and Technology, ICIMTech 2020, pages 549–554. Export Date: 24 November 2021.

Mitchell, T. M. (1997). Machine learning.

Mudde, C. (2000). *The Ideology of the Extreme Right*.

Necaise, A., Williams, A., Vrzakova, H., and Amon, M. J. (2021). Regularity versus novelty of users' multimodal comment patterns and dynamics as markers of social media radicalization. In *HT 2021 - Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, pages 237–243. Export Date: 22 November 2021.

Neumann, P. R. (2017). *Radicalized: New Jihadists and the Threat to the West*. I. B. Tauris & Company, Limited, London.

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. Cited By :14 Export Date: 22 November 2021.

Oussalah, M., Faroughian, F., and Kostakos, P. (2018). On detecting online radicalization using natural language processing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11315 LNCS, pages 21–27. Cited By :5 Export Date: 22 November 2021.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ramirez, T. A. F. (2021). Identifying characteristics of persons vulnerable to social media extremism.

Rekik, A., Ameur, H., Abid, A., Mbarek, A., Kardamine, W., Jamoussi, S., and Hamadou, A. B. (2018). Building an arabic social corpus for dangerous profile extraction on social networks. *Computación y Sistemas*, 22.

Rekik, A., Jamoussi, S., and Hamadou, A. B. (2020). A recursive methodology for radical communities' detection on social networks. In *Procedia Computer Science*, volume 176, pages 2010–2019. Export Date: 22 November 2021.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature (London)*, 323(6088):533–536.

Russell, S. and Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Prentice Hall series in artificial intelligence. Pearson Education, Limited, Harlow.

Saif, H., Dickinson, T., Kastler, L., Fernandez, M., and Alani, H. (2017). A semantic graph-based approach for radicalisation detection on social media. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10249 LNCS, pages 571–587. Cited By :19 Export Date: 22 November 2021.

Saif, H., Fernandez, M., Rowe, M., and Alani, H. (2016). On the role of semantics for detecting pro-isis stances on social media. In *CEUR Workshop Proceedings*, volume 1690. Cited By :2 Export Date: 22 November 2021.

Skagseth, A. E. N. (2021). Bts army: En studie av buying parties som en sosioteknisk praksis i en k-pop fandom.

Sutton, R. S. (2018). Reinforcement learning : an introduction.

Trevor, v. M. (2014). The 1% rule in four digital health social networks: An observational study. *Journal of Medical Internet Research*, 16(2). Copyright - © 2014. This work is licensed under http://creativecommons.org/licenses/by/2.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2021-04-16.

Tuters, M. and Hagen, S. (2020). (((they))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*, 22(12):2218–2237.

Ul Rehman, Z., 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. Cited By :3 Export Date: 24 November 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Wang, Y., Liu, T., Tan, Q., Shi, J., and Guo, L. (2016). Identifying users across different sites using usernames. *Procedia Computer Science*, 80:376–385. International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA.

Ware, J. (2020). *Testament to Murder: The Violent Far-Right's Increasing Use of Terrorist Manifestos*. ICCT Policy Brief.

Yadav, S. and Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 78–83.

Zhang, H. and Su, J. (2008). Naive bayes for optimal ranking. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(2):79–93.

Appendices

A Quality Assessment Criteria

All the quality criterias are heavily inspired by quality criterias from the paper *"How to do a Structured Literature Review in computer science"* by Anders Kofod-Petersen [Kofod-Petersen, 2015].

QC 1 Is there is a clear statement of the aim of the research?

QC 2 Is the study is put into context of other studies and research?

QC 3 Is the decisions of approach or use of algorithmic justified?

QC 4 Is the algorithm reproducible?

QC 5 Is the test data set reproducible?

QC 6 Is the study approach/algorithm reproducible?

QC 7 Is the experimental procedural explained in details?

QC 8 Is the results evaluated with metrics?

QC 9 Is the results discussed?

QC 10 Do the findings get supported by the evidence from the test?

QC 10 Does the paper propose an approach with NLP?

B Extracted Data from SLR Papers

The following information was extracted from the literature study. The information is presented in Table 1 and Table 2.

- ID
- Title
- Author(s)
- Year of publication
- Algorithm used to identify traits or indicators of vulnerable people for radicalization
- Algorithms or elements used from in NLP
- Type of Extremism

- Selected features used
- Data-set
- Data-set origin
- Relevant findings/Conclusions
- Prediction/Identification

C Classifiers' Parameters in Experiment 1 and 5

DecisionTreeClassifier
criterion: default="entropy", used with Shannon information gain.
splitter: default="best", means it choose the best split.
max_depth: default=None.
min_samples_split: default=2.
min_samples_leaf: default=1.
min_weight_fraction_leaf: default=0.0.
max_features: default=None, means max_features=n_features.
random_state: default=None.
max_leaf_nodes: default=None, means unlimited number of leaf nodes.
min_impurity_decrease: default=0.0
class_weight: default=None, means all classes have weight one.
ccp_alpha: default=0.0.

Naïve Bayes:
priors: default=None, means no prior probabilities of the classes.
var_smoothingf: default=1e-9

Logistic Regression:
penalty: default='l2', means it add a L1 penalty term.
dual: default=False, means dual or primal formulation.
tol: default=1e-4, referees to tolerance for stopping criteria.
C: default=1.0, is a value that inverse of regularization strength.
fit_intercept: default=True, means if a constant/bias should be added to the decision function.
intercept_scaling: default=1.
class_weight: default=None, mean all classes are supposed to have weight one.
random_state: default=None, important if using "sag" as solver. *solver*: default='lbfgs', means which algorithm to use in the optimization problem.
max_iter: default=100, means the maximum number of iterations taken for

the solvers to converge.
multi_class: default='auto', means 'auto' selects 'ovr' if the data is binary, or if solver='liblinear', and otherwise selects 'multinomial'.
verbose: default=0, means the level of verbose.
warm_start: default=False, means it erases the previous solution.
n_jobs: default=None, means the number of paralleling processes used if parallelized.
l1_ratio: default=None.

D Tables from Step 5 in SLR

ID	Title	Author(s)	Year	Algorithm(s)	NLP Element(s)	Extremism	Features	Dataset	Origin	Findings/Conclusion	D/P
1	Using KNN and SVM based one-class classifier for detecting online radicalization on twitter	Agarwal and Sureka	2015	SVM, KNN, Semi-supervised Learning	TF	Jihadism	TF vectors	UDI-TwitterCrawl-Aug2012 ATM-TwitterCrawl-Aug2013	Twitter	Internet slangs, emoticons, and punctuation are less important features in KNN. Not the same for SVM. C: Presents of religion, war-related terms, offensive words, and negative emotions are strong indicators	D
2	Countering Terrorism Incitement of Twitter Profiles in Arabic-Context	Alghofaili and Almishari	2018	Random Forrst, Naive Bayes, J48, SVM, Naive Bayes Multinomial	Classic preprossesing, Feature set-vector from number of words used by user(entire feed)	Terrorism Jihadism	TF vector for each user feed.	600 accounts(T= 3200) where 100 were terror incentives profiles.	Twitter	Classical ML methods can return up to 85% accuracy	D
3	Intelligent detection of hate speech in Arabic social network: A machine learning approach	Aljarah et al.	2021	SVM, NB DT, RF	TF, TF-IDF, BoW	Hate Speech	TF TF-IDF BoW	Self-labeled as positive or negatie case of hate speech.	Twitter	Good results were archived with ML approaches Best being RF with TF-IDF (accuracy=0.882, G-mean=0.882, P=0.9)	D
4	Detection of Violent Extremists in Social Media	Alvari et al.	2019	SVM, Char-LSTM LabelSpreading Laplacian SVM Vo-training KNN Gaussian NB Logistic Regression AdaBoost Random Forest	(?)Levenstain Distance	Jihad	Username	1.6 M tweets based on 25 extremism hasthagss	Twitter	Username can contain valueavle data to idetify radical people returning to twitter. Good results only by using twitter username.	D/P (With user-name)
5	Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter	Benigni et al.	2017	Multiplex vertex classification, MNVC	No NLP	Jihad and Online Extremism Commuinty	Graph-theory	Seed users followers and what they follows. Snowballing	Twitter	Interesting approche to creating dataset	-
6	User Identity Linkage in Social Media Using Linguistic and Social Interaction Features	Chatzakou et al.	2020	Naive Bayes, BayesNet, J48, LADTree, LMT, Random Forest (Esemble Method) RNN, GRU,	Characters-based -, Word-based -, Sentence-based -, Dictonary-based frequency. POS tags, Word2Vec Levenstain Distance	Terroism, Abusive language (hate-speech)	Characters-based -, Word-based -, Sentence-based -, Dictonary-based frequency. POS tags, Word2Vec Levenstain Distance	It consisted of two datasets, "The abusive Dataset" and the "Terrorist Dataset", where the first is from Chatzakou et al. [2017], while the other was created with term searching terrorist-related keywords(65k).	Twitter	The NN had the most impact by these features in contrary to activity and network features. Random-forest showed the best results based on all the features, and they concluded that text similarity improved the results of all cases. Results of ROC 99.5%	D(?)
7	A Study on Keyword Analytics as a Precursor to Machine Learning to Evaluate Radicalisation on Social Media	Chelvachandran and Jahankhani	2019	Non	Word-counting	Terrorism, Jihadism	Word Frequency	1. ISIS English-based magazine, - Dabiq(15 issues) - Rumiyah(9 issues) Baseline for keywords(2685 text) 2. 17'000 tweets from 100+ pro-ISIS profiles Both from the US Deparment of Defence research contractors.	Twitter	Keyword usage can be used to identify radicalization and exploitation in social media. However, a depth-analyses is needed on the words. Difficulties around Twitter are deleting users.	-
8	A Framework for Predicting and Identifying Radicalization and Civil Unrest Oriented Threats from WhatsApp Group	Deb et al.	2020	Semi-superviser learning, Mining: FP growth Algorithm, SPMining NSPMining	Word Stemming Pos-tagging	Not Defined - Terrorism - Poltical - Rioting	Stemming(Porter) POS-tags Pattern regonition - Frequency W - Pattern W - MScore	The approach is based(?) on the inference of chat logs as the selection of the enemy, negative emotions toward the enemy, and promotion to other protesters	WhatsApp	Multimedia shows to be a part of radicalization as well. Repetition of specific phrases also.	D/P
9	Eventfully Safapp: hybrid approach to event detection for social media mining	Derbas et al.	2020	Used, but not defined	Synonyms: WordNet Syntactic extraction based on rules	Detcted Events	Regex rules, Rule-based extraction	The dataset was sentences created by the author(s) by annotating by hand. It contained 200 sentences of events(kill events)	Twitter/ Safapp	Tried to modify a solution called Safapp developed by EU, to find "Kill" events. Small dataset. Results were good, but room to improve. Could improve ontology of the word of radical and to find the social sematice value on topic after events.	D(?)
10	Contextual semantics for radicalisation detection on Twitter	Fernandez and Alani	2018	SVM, Naive Bayes C, DT, J48	Sematic analysis Context-SA(TextRazor)	Terrorism Jihadism ISIS/ISIL	NER N-grams	114k tweets dataset: 17K posted by pro-isis users 97k posted by general users "How ISIS uses twitter"from Kaggle where they "approved pro-ISIS users" by checking if they are blocked on Twitter now. General counterpoint the same by been active for 2 years.	Twitter	Usage of the semantic context of terms that are linked to radical rhetoric improves the detect radical content. The SVM with sematic based radicalization detection (P=0.859, R=0.843, F=0.851) vs not (P=0.816, R=0.801, F=0.822)	D/ (Could become P)

Table 1: SLR overview 1/3

ID	Title	Author(s)	Year	Algorithm(s)	NLP Element(s)	Extremism	Features	Dataset	Origin	Findings/Conslutions	D/P
11	Understanding the roots of radicalization on twitter	Fernandez et al.	2018	J48, Naive Bayes, Logistic Regression	Word-frequency N-grams Pre-prssesing - Numeric removal - Punctation removal - Stopword(Ranks) - URLs	ISIS Terrorism Jihadism	Root of radicalization micro, meso and macro N-grams(in RoR)	Terms: ICT Glossary Saffron Experts Saffron Dabiq Magzine Rowe and Saif Dataset: How-ISIS-Uses-Twitter Datast-Spotlight Isis-Realted-Tweets	Twitter	Proposes a computation approach for detecting and predicting the radicalization influence on users based on the social acience model "Root of radicalization". The approach is tested against 112 pro-isis users and 112 general users from Twitter. Results show an 0.9 F-1 score for detection and 0.7-0.8 in prediction.	D + P
12	Predicting online extremism, content adopters, and interaction reciprocity	Ferrara et al.	2016	Logistic Regrssion Random Forest Evaluation: Cross-validation	Minimal,contained features frequency	ISIS	52 Twitter user features about Users metadata and activity, Timing-data, and network data.	Lucky Troll Club"s 25k pro-isis users. Retrived with Twitter API	Twitter	Conclusion that the ratio of retweets, average number of hashtags, number of tweets and average retweet per user are high in discriminative power for prediction	D + P
13	Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit	Grover and Mark	2019	Non, Statistic approche	TF-IDF, Word-Frequency Terms/sematics: LICW HateSonar	Alt-Right	F-testing and more statistics(?)	r/altright (6 months)	Reddit	By using TF and TF-IDF identified the frequent terms that corresponded with racial slurs and anti-Semitic words. Used theory to calculate with statistics two traits, fixation and group identity. Found increment in both traits from the dataset of comments	D
14	Identifying right-wing extremism in german twitter profiles: A classification approach	Hartung et al.	2017	SVM	BoW, Bi-gram	Rigth-wing	Lexical, Emotions Pro/Con features Sosical Identity Featurs	Seed profiles of Twitter users (n=37, RW=20, N=17) Entire user feed	Twitter	(?)	D
15	Meta-terrorism: Identifying linguistic patterns in public discourse after an attack	Kostakos et al.	2018	Passive-aggressive classifier	Sentiment Analysis - SentiWordNet Topic Analysis Fake News Detection	Terrorism - Manchester - Las Vegas	Not clear	Gathers data from terror events with API. Uses it to nalayse sentiments, topics and detect fake news.	Twitter	Dataset extraction is a bit unclear, but it looks like they use the API to extract themself. Sentiment aanlysis shows clear spike	A/D
16	Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter	Lara-Cabrera et al.	2019	Other field: Social Network Analysis	Preprossesing: - Regex and Removal	Jihad	Word-Frequency Analysis	D1:"How-ISIS-uses-Twitter" D2: Anonymus OpISIS D3: Random	Twitter	LaraCabera mentions that frustration is a trait that can be measured in how many swear words and the format of the sentences and that introversion is measured in the length of the sentences(short).	P
17	Towards the automatic identification and monitoring of radicalization activities in twitter	López-Sáncez et al.	2018	N/A Used mathematical appoche of two equations.	N/A	Far-Right "Hogar Social Madrid"	Users Meta-data	API usage to detect tweets, fufilling the search term (Logical operators) to the Twitter API. Downloaded with meta-data of user also.	Twitter	Found radical users and users with the potential of being radicalized. The approach found data with terms and the help of an expert and then used the meta-data to find users. Found the users with early signs of radicalization. FW: Combine with NLP :)	D/P
18	A study of radicalism contents detection in twitter: Insights from support vector machine technique	Miranda et al.	2020	SVM	Preprossesing TF-IDF	ISIS	TF-IDF	Retrived Tweets with API. 100tweets where they where hand-labeled(?) 61 Radical and 39 not.	Twitter	The paper is more research on the usage of a classic approachs to use the method on the Indonesian language. Have not much to contribute other than that it is not so much research done on the Indonesian language and is some sort of an initiator of creating a baseline.	D
19	Regularity Versus Novelty of Users' Multimodal Comment Patterns and Dynamics as Markers of Social Media Radicalization	Necaise et al.	2021	N/A	Nonlinear Dynamical system theory(NDST)	(?)	Meta-data from user for two-years - Ratio on comments in R subreddits or not - Length of comment - score(upvote)	The dataset was labeled based on the frequency of published posts on radical subreddits.	Reddit	Concluded with that highly active users (multiple radical subreddits) are more novel in commenting over time, meaning more changes than other medium-low ranking radical users. Potential for good preditors. It was more of an analytic paper than a framework.	P/A
20	Understanding the radical mind: Identifying signals to detect extremist content on Twitter	Nouh et al.	2019	SVM KNN NN RF	TF-IDF N-gram(1,2,3) Word2Vec(Skip) LIWC Dictonary	ISIS	Diffrent parameters for: Radical Language and, Psycological signals	Two methods:1) Create a Word Embbeding from Dabiq - TF.IDF -Word2Vec 2) Create a psycological profile Used on : "How-ISIS-Uses-Twitter"	Twitter	Compares the methods of KNN, SVM, NN, and RF(is mentioned before as best) with the words. The finding shows that potential TF-IDF is not as good performing as word2vec, and NN, RF showed bed results Findings found that Us-terms appare often, more hashtags and radical psyco distance in radical users.	D

Table 2: SLR overview 2/3

ID	Title	Author(s)	Year	Algorithm(s)	NLP Element(s)	Extremism	Features	Dataset	Origin	Findings/Conslutions	D/P
21	On detecting online radicalization using natural language processing	Oussalah et al.	2018	KNN-SVM RF	(Dataset) LIWC MRC-PsycolinguisticDB WordNet	(?)	Meta-features Sematic Analysos	Created based on hastags: Twitter: 12.220 tweets revied by Amazon Mechanical Turk Tumblr:	Twitter Tumblr	Two methods based on a metric proposed in the paper, while the second uses a KNN-SVM machine learning model. The ML approach includes N-grams, personality traits, emotions, as well linguistics, and network-related features. The conclusion is that the metric can be helping the system improve.	D
22	A recursive methodology for radical communities' detection on social networks	Rekik et al.	2020	N/A	N-gram Preprossesing	Jihad	Non, was callculated with Cohen's Kappa coefficient(?)	They collected radical data from Youtube and Twitter based on the methodology presented in Rekik et al. [2018]. Created a dictonary with n-grams that updates/change over time by retiving new radical users.	Twitter Youtube	It created a system rooted in data mining that collects vocabulary by using n-grams and proposed math equations to find the "violence in the word". The system also updates/changes since it is a recursive system. Promesing results.	D
23	A semantic graph-based approach for radicalisation detection on social media	Saif et al.	2017	SVM (kernal=RBF) MaxEntropy NaiveBayes	NER DBpedia	ISIS	Graph-based features Unigrams Sentiment Feature - SentiStrength Topic Feature - LDA Network Features - Meta-data from user	Previous work dataset [14] (in paper) contains 1132 european Twitter users, where 727 are pro-ISIS based on the user's content.	Twitter	Best results with SVM where the sematic results were used had (R=+7.8% P=+7.7% F1= +7.89%) better than the average of the rest using the other features. Sematic features are suitable for classifying the user for or against ISIS.	D/A
24	On the role of semantics for detecting pro-ISIS stances on social media	Saif et al.	2016	Naive Bayes	Bag-of-Word Unigram	ISIS	Sematic features Network featurs BoW Feature analyse with Information Gain	Dataset form previous work[4] (in paper): 1132 Twitter users - 566 Pro - 566 Against	Twitter	The semantic features are extracted with AlchemyAPI from DBpedia, YAGO, OpenCyc, Freebase, etc. Sematic Features improved the system by 2% (F1-score)	D/A
25	Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning	Ul Rehman et al.	2020	Naive Bayes SVM Random Forest	Preprossesing - Tokenization - Normalization - URL removed - Stop-word removal - Stemming Plus: - Removal of non-english terms Detection of terms: TF-IDF N-gram	ISIS	(?)	D1:Radical corpus "How-ISIS-uses-Twitter" D2: Neutral corpus The anti-ISIS in "ISIS-Related-tweets" D3: Religious corpus Dabiq and Rumiyah magazine "isis-religious texts" D4:New Dataset Crawling on the account Ctrl-Sec called out for radical befor suspended D5: Random Dataset 7000 tweets manually check to be non-isis related.	Twitter	Religious terms in the dataset were detected using the highest scoring after TF-IDF with N-grams; as shown in other work, this method efficiently detects hateful and extreme content.	D

Table 3: SLR overview 3/3

Thomas Alejandro Fernandez Ramirez

How to Catch a Far-Right Radical

NTNU
Norwegian University of
Science and Technology