Paul Verhaar 3641309

Supervisor: Mirko Tobias Schäfer

Second reader: Stefan Werning

June 2016, Utrecht University

Master New Media & Digital Culture

Radical Reddits: into the Minds of Online Radicalised Communities.

Towards a classifier for recognising radicalised language coordination within the Reddit forum.

Abstract

The online domain potentially provides research with a vast body of data. The big quest within contemporary research is to make sense of all this data. A possibility to handle such data is by combining methods from the fields of new media and linguistics. Several studies have sought to understand how radicalism online comes to exist and grows. To date, however, none of these studies have fruitfully analysed language patterns within online communities that move beyond keyword analysis. In this thesis, I demonstrate a proof of concept for a classifier analysing and predicting salient language features within online radical discourse from the social media platform Reddit. Data consist of two datasets, radical and non-radical in nature, both containing 1 millions lines of text per dataset. The radical dataset is known for its radical nature, promoting radicalism in a variety of beliefs such as anti-feminism or white supremacy. Using software libraries as NLTK and SciKit within Python, I submitted that data to keyword and collocation frequency count, lexical diversity, a part-of-speech tagger and ultimately as features for a document classifier. Results showed that the radical discourse used in this thesis contains salient language features and show a clear sense of a virtual community. Finally, I discuss the implications of this thesis and provide directions for further research. Data was provided by TNO The Hague as part of the VOX-Pol project.

Keywords: online radicalism, Reddit, virtual communities, computational linguistics, data science.

Foreword

The adventure of this master's degree started off by taking some courses to broaden my knowledge. Little did I know I would be fully submerged in and enchanted by performing research within the fields of new media and linguistics combined. The path that this degree, as well as this thesis had led me to is one that suits me well. This academic year has shown me where my focus should be: doing meaningful research within the online realm by combining theories and techniques from both studies.

As with all theses, it is never an easy victory. Throughout my years of studies, I have learned that there is no victory without effort, or hard work in this case. I have been very fortunate to find the right supervisors and people around me to motivate me in doing research and learn more within the field of new media every day. And, as with all research, it is a process of trial and error. For this thesis, I must say, it was an intense ride; early mornings, or shall I label them as late nights. Perhaps I lost track. In any case, I did learn to become semi-professional in Googling. You could say that this master's track paid of. The biggest challenge for this thesis was to touch on slightly new or at least scarce fields of research while maintaining a focus on new media. But, thanks to my supervisor at the University of Utrecht and TNO, I managed to make it to the end.

Upon writing this part, I can truly say that this thesis is done. This means that I have finished my master's degree in New Media & Digital Culture. It has been a hell of ride, which I would not want to change for anything else. Finally, a big applause for my friends and family. Without them I would not have made it this far and have finished this thesis. You know who you are. Some heroes don't wear capes.

Table of Contents

1. Introduction

Internet and language play a ubiquitous role in our lives and the way we communicate. They both serve as a powerful political tool that affords the imminent threat of online radicalism in Western contemporary society. Radicalised communities are more capable than ever to spread their ideologies through online communication technology as websites and more recent, social media. On the other hand, social media platforms do not want to facilitate this as is shown by the social networking site Reddit that banned the Coontown community for containing radicalised posts (Newitz 2015). In conveying radicalised beliefs, language plays an important role within the field of sociolinguistics. This has been widely researched offline (Barton and Lee 2013). However, the amount of research for online sources remains scarce (Schwartz et al. 2013). The Internet as a communicative tool affords online radicalisation, resulting in vast pool of data (von Behr et al. 2013). Recent studies looking into the effect of the internet on radicalism have shown that non-US based extremists were more likely to learn through virtual tools (Gill, Corner & Thornton 2015). The web enhances possibilities for radicalised groups to communicate, organise and plan activities (Easttom & Taylor 2011); they use the power and the freedom of social media platforms to exercise pressure, power and influence across the globe (Graham 2013). They post messages to support their radicalised beliefs in terms of views, sermons and propaganda for like-minded users. As proposed by Paßmann, Boeschoten and Schäfer (2014) within the new media discipline, there are several reasons that results in the use of such platforms. Examples of this are found in (self)-profiling and sharing common values which result in online relationships between users.

Radicalisation not only poses a threat, but also creates the necessity for further research within social media platforms. The earliest piece of analysis on violent radicalism and the Internet appeared in 1985, but the vast bulk only began to be produced in the 2000s, with a significant uptick since 2010. Because of this, older methods need to be adapted, replaced or combined with new methods that allow for contemporary research (Rogers 2013). Methods for this are labelled as digital methods, wherein the nature of methodology lies in the epistemology (Rogers 2013). The use of digital methods will allow researchers to use social media platforms as a data source that informs on social processes (Passman et al., 2014). This allows researchers and institutions to seek out specific patterns of language coordination, social ties and posting behaviour.

Throughout recent years, researchers have been fascinated by the role of language in social media (Conover et al. 2011, Castillo & Poblete 2011, Danescu-Niculescu-Mizil et al. 2012). It has mainly focussed on extracting social network graphs from a collection of social media messages, which rely on surface statistics, as message frequency, followers and reciprocity (Castillo & Poblete 2011). Some conducted research on network structures without paying attention to the linguistic patterns in their set (Conover et al. 2011). Unfortunately, only small results have been found for indicators of language patterns online such as hate speech (Balcerzak

& Jaworski 2015). What is more, they mostly neglect deeper linguistic analysis of content, keyword abstraction and language structures. Research on radicalism and language has had a main focus on the qualitative nature of language analysis, with a scarce body of research focussing on the online realm (Brindle 2009 & Duffy 2003). Language research online mainly targeted forums and chat rooms (Sproat et al. 2001; Aw et al. 2006; Han and Baldwin 2011; Yang and Eisenstein 2011); while none of them reached firm conclusions. Also, research has primarily looked at comparisons between different websites or older radicalised forums such as Stormfront[1] (Koster & Houtman 2016). Despite a focus in some areas focus, there is still a lack of interdisciplinary research (Conway 2016). It seems that there is a growing need for corpus linguistic techniques in the field of new media to bridge the gap between media research and radicalised communities online. This thesis considers how automated corpus linguistic techniques may be used to facilitate the process of identifying the ideology expressed in radicalised language material.

The current work explores the possibilities of combining linguistics and new media methods for exposing salient language features in online radicalism. The goal of this thesis is to dissect language patterns from online radicalised communities to train a model that can analyse language from online forum posts on the content of radicalised language coordination. This will be performed by using a machine learning algorithm in which a computer predicts whether or not an online text is radical in nature. Mapping radicalised language patterns can help us understand their language behaviour. Another important aspect of the current study is that it incorporates an up-to-date source as Reddit, which is a publicly approachable moderated forum with an abundance of posts. This source was chosen since Stormfront is not seen as the most violent forum anymore, but is caught up by Reddit (Hankes 2015). Reddit boasts the 9th highest Alexa Internet traffic ranking in the United States and the 36th worldwide. Many of Reddit's radical sub-reddits are among its most popular (Hankes 2015). The ever-growing structure of this forum as well as its status amongst youngster allows the current research to establish a solid data pool. The analysis presented in this work addresses previous shortcomings by employing a quantitative and qualitative approach including keyword analysis, frequency counts, parts of speech tagging, lexical diversity and a proof of concept for document classification to examine salient language use within radicalised communities. This work will do so guided by the following research question: which distinctive patterns of language can be found for language coordination within radicalised online forum communities? In order to answer the main question, this paper addresses the following sub questions: What is the frequency distribution in the dataset? What are the most common keywords within the dataset? What are the co-occurrence patterns for words? How do they compare to other posts within non-radicalised sections of the forum? The project has been conducted as part of the VOX-Pol project at the Dutch research institute TNO in the Hague[2]. Through this exploration, the

[1] https://www.stormfront.org/forum/
[2] http://www.voxpol.eu/ / www.tno.nl

goal is to gain insight into the language use of radicalised communities and test the effectiveness of corpus linguistics for such purposes.

An analysis as well as a theoretical foundation is needed to answer the research question. Chapter 2 will review existing literature and its results in light of the current work. Subsequently, chapter 3 will give an overview of the used method of analysis. Chapter 4 contains a description of the data and the results obtained in this study. The current work and its implication will be discussed and reflected on in chapter 5 as well as provide suggestions for further research. Lastly, chapter 6 will conclude this work and provide a summary of the findings.

2. Theoretical Framework

Sections 2.1 and 2.2 touches on the definition of radicalism, virtual communities, and the role of radicalism in those within the field of new media. Section 2.3 reviews existing theories and literature from the field of linguistics. In this, it will include linguistic phenomena as register, homophily as well as previous linguistic studies on radicalism online. Section 2.4 reviews literature from the field of machine learning and document classification.

2.1 Defining radicalism

The term ‘radicalisation’ is widely used online, but in general lack a universal definition. Even more so, the definition of radicalisation has been a topic of recent debate Coolsaet (2011). The course of history shows that the term radicalisation can be used in a vast array of circumstances; dating back to political parties from the 19th century where the term was used to signal a flow of change within the political department and was sometimes even referred to as being a non-violent activist (Schmid 2013). When putting this next to the contemporary use of the word, it points at a different kind of activism that is anti-liberal and fundamentalist. For example, Coolsaet (2011) states that the definitions as currently displayed ‘ill-defined, complex and controversial’. Coolsaet (2011) and the course of history show that radical is a relative term as well as a hard term to describe or define. The term itself is an ongoing term that constantly changed and has been adapted to its times and spaces in which it is used. Coolsaet (2011) therefore argues that radicalisation is a process, without giving a clear definition of the word itself. In relation to this, radicalisation is most commonly used in relation to Jihadism or Islamic radicalisation (Hoskins 2011). However, other notions of radicalisation are also seen as part of Fascism or white supremacy (Bowman-Grieve 2009). As Conway (2016) correctly argues, (violent) radicalisation is not something that should only be investigated in the context of Jihadism; while important it should not be the sole consideration of research and the term should be put to use more broadly.

For the current thesis, the definition of radicalisation will be used in that it defines acts or

political activities against a mainstream democratic society (Sedgwick 2010). I expand it by adding work by Midlarsky (2011). He defines radicalism as a will to power in which a social movements and its acts thereof are the vehicle to get to power. In this, radicals “strive to create a homogeneous society based on rigid, dogmatic ideological tenets; they seek to make society conformist by suppressing all opposition and subjugating minorities” (Midlarsky 2011). Online radicalisation can be seen as a process wherein individuals through their online interactions an exposure to various types of social media content, come to view violence as a legitimate method of solving social and / or political conflicts.

2.2 Radicalism online: virtual communities

Over the past years, an average of 15% of the world population has expressed themselves through language on the wide variety of social media channels available (Hauffa et al. 2014). Social networking sites are a platform that make it possible for its users to interact (Anger & Kittl 2011). Social media platforms provide groups of all sorts with a powerful tool for information provision and the opportunity to share beliefs and ideas on a global scale, as well as using it for psychological warfare (Conway 2002). It allows radicals to use intra- and inter-group communication to transform and build new structures for information spread. As Margulies (2004) proposed, social media provides radicalised groups with a greater online communicative efficiency. It does not only connect affiliated members within one group with each other, but also breeds connections between different groups that might have been unaware of each other before. This results in a vast amount of information spread for which social media platforms are perfectly suited. The easy access nature of them is what makes these platforms such a success, according to Zhao and Rosson (2009). Other research, amongst which Passman, Boeschoten and Schäfer (2013), claim that participation on online platforms and the co-creation is an added value because it affords profiling, sharing shared values and maintaining online relations by showing mutual appreciation between users. Rogers (2004) argues that social networking platforms are “walled-off echo chambers” that resolve around their own established views. It shows that information cannot be taken as it is and has to be evaluated correctly.

2.2.1 The ‘we’ sense

The use of media and interaction affords the sense of virtual community (Rheingold 2000). A virtual community in this surpasses the notion of physical community since its users do not have to be physically present to create the we-sense (Anderson 1991). Users on the Reddit forum can be placed in such a category, wherein their active participation on the website within sub-reddits plays an important role for community building. The construction of virtual communities rises and falls with the use of media since online presence gives users a sense of being part of a non-physical

community. The use of sub-reddits is an example of how posts on certain topics attract different users in which involvement within that topic creates a community feeling. Rheingold (2000) argues that all information that is spread contributes to the shaping of communities. Such communities, if big enough, might create the sense of "we" within the community (Rheingold 2000). Radicalised communities can be viewed as such a community, with shared values and a sense of cohesion. O'hara and Stevens (2015) claim that social media platforms are also a sharing tool for views that depart from society's standard in which prejudices might be reinforced. Connected to their view, users online position themselves within virtual communities wherein similar messages or beliefs are spread; such communities are also referred to as "echo chambers" (O'hara & Stevens 2015). This concept is not new, and has been widely applied to the web and online communities in which it breeds new social patterns (Lanier 2011). Paragraph 2.3 builds on communities in a language sense.

2.2.2 The radicalised community

Radical movements are known for their creative and innovative use of virtual communities. Their use has not been limited to forums or bulletin boards but expand to other virtual communities such as social media platforms. This provides us with a rich pool of information for analysis. Previous research has shown that members of virtual communities associate with each other based on mutual interests (Bowman-Grieve 2009). As stated before, radicalism does not restrain itself to Islamic extremism or terrorism, but is used in a broader sense throughout this thesis (Midlarsky 2011). In light of this definition, it is important to look at research that also covers other radicalised groups, such as white supremacy, right-wing extremism, fascistic parties or fundamentalist Christianity. An example of such a group is the Stormfront community, wherein Bowman-Grieve (2009) has found that active participation is key for maintaining and letting a community grow. A study by Zhou et al. (2006) also analysed Stormfront which, despite its age, is still a relevant forum for analysis. It is among one of the first white supremacy "hate sites" (Whine 1997).

Many scholars hold the view that echo chambers are polarising, and are afraid that this will lead to violent acts (O'Hara & Stevens 2015, amongst others). The concept of the echo chamber, however, has been mainly dealt with in the sense of Islamic extremism, and has been made priority for policy intervention (Sunstein 2007). Echo chambers facilitate social fragmentation, wherein diversity is a marker for polarisation. Despite the fact that online spaces are mostly public spaces and subjective to free speech, they are extremely important to monitor communities and their behaviour (Sunstein 2007). Sunstein (2007) goes on by arguing that the Internet is not a public forum, since many information is tailored to the users' specific needs; members are mainly exposed to arguments in favour of the community or group. Echo chambers have even resulted in blogs having become an important part of radical networking (Bunt 2009). On the other hand, echo

chambers can also interact with ideological opponents, as shown by Gruzd & Roy (2014) who found that users on Twitter engaged opponents, but only did so to portray their standards. To facilitate a change in point of view, the links towards the outside of the echo chambers have to be convincing and solid. In echo chambers, there are multiple factors that act as the glue, but all of them are part of the homophily of a group. In this, groups gradually become more homogenous over time which also involves becoming more radical as the consensus within the group grows (O'Hara & Stevens 2015).

Despite the fact that research on radicalised groups online has gained attention, it is still in its early stage (Burris, Smith & Strahm 2003; Gustavson & Sherkat 2004; Boutyline & Willer 2015; Chen 2007). One of the biggest points of critique is that the majority of the studies into radicalism online have not reached firm conclusions. It lacks current value for contemporary debates and the shift to more popular social media platforms such as Reddit.

2.3 Internet language

Language variation is a ubiquitous part of communication, and is more than ever evident in new forms of writing such as social media or blogging. Language online has been a broad topic of investigation in which language used for online communication has been named "internet language" (Crystal 2001). While some have proposed one single variant of online language use, reality has shown differently. Thurlow (2006) has shown that many of the linguistic coordination found online already existed before the upcoming of new technological tools. This shows that not only language itself, but also online practices are an important topic of research. Important research within this field has been carried out by Danescu-Niculescu-Mizil et al. (2011); he found that amongst Wikipedia users, like-minded users share linguistics features within the community. Similar research has been backed by theoretical foundations as homophily, register and social stratifications, which will be explained in the following sections.

2.3.1 Communities: linguistic phenomena

Language analysis online has been carried in the past with a focus on email analysis (Groh and Hauffa et al. 2011), analysis on retweet behaviour (Passman et al. 2014), abstracting personality from social media (Schwartz et al. 2013), style accommodation (Danescu-Niculescu-Mizil et al. 2011) and extremist forums (Attestog & Perera 2013). Many concepts are interchangeable between the fields of new media and linguistics, such as homophily, accommodation theory and register.

Homophily is a concept that is part of social media studies as well as linguistic studies. Homophily is a principle about connection, that can be structures from a network of people or language structures (Biswas 2016). It results in behaviour of people that is connected to a certain

network with “sociodemographic and interpersonal characteristics” (McPherson et al. 2001:415). Their network and relations have a powerful impact on not only the information and interactions they receive, but also on the information and interactions they (re)produce. Homophily in a linguistic sense not only influences people’s social worlds, but also their language: people tend to use similar language within a certain group or setting, just as they form similar attitudes within that group (Gilbert 2012). People adapt, shape and are influenced by their environment in terms of networks and language (Jasnow et al. 1988).

The theory of homophily is closely related to a predominant theory in linguistics by Giles (1991) called the accommodation theory and is positioned as an “integrated, interdisciplinary statement of relational processes in communicative interaction” (Giles 1991). The theory looks at predominant homogenous language within the group and less at the relations of users within the group expressed via language. Thus, social network homophily and accommodation theory have the potential to provide a general and effective way to account for linguistic variation in natural language processing online.

Another important theory for this thesis is register, which is an important part of our everyday conversation, such as during face-to-face interaction, but also occurs in the online domain (Biber & Conrad 2009). With register, people tend to convey their messages in different ways and forms depending on the person they share information with (Danescu-Niculescu-Mizil et al. 2011). Taken more broadly, language homophily is connected to register, i.e. language use that shapes information in specific occasions and shapes the network of the person itself. Register is a strong and important aspect in this, as proposed by Pennebaker (2011), who claims that words can be a “window to the soul”. In other words, the register used can signal a specific use for a community and the people inside it. The linguistic characters that are used can be seen as markers within that group and can be shown as being a target register. Danescu-Niculescu-Mizil et al. (2012) showed the importance of such a target register in his study on how people accommodate on Wikipedia. They proposed that language coordination is strongly dependent on the accommodation theory and power differences within social groups. Their interactions rely on linguistic style markers, such as the use of content and function words and certain keywords (Anger 2011).

Register, consisting of linguistic style markers, contains the building blocks for unfolding and characterising a variety of language. By using language, people divide themselves and others into different groups by taking class, status and power into account (Nichols 1984). Social stratifications play an important role not in the use of language, but also in the perception of language. The linguistic behaviour as part of groups is an indicator of social presence within a community. The experience a speaker, or user, has is mediated through the experience of text on social occasions (Kress 1989). Despite language having “different look” to different users, it is possible to discern similar linguistic markers between groups (Pennebaker 2011). For example,

function words serve a grammatical role in a sentence and can therefore be seen as a linguistic marker for attitude or mood (Pennebaker 2011). Pronouns tell us where people focus their attention and who is of higher status (Pennebaker 2011). Summarising, not only the network or community a person moves in, but also the language is important for analysing online social formations on social media platforms.

2.3.2 Language use: radicalism

Language is seen as the most important tool for classifying human diversity (Hutton 1999). To date, research on Internet language within online radical communities remains scarce. This, while radicalised communities "across different jurisdictions heavily utilise modern transportation and communication systems for relocation, propaganda, recruitment, and communication purposes" (Chen et al. 2004).

Research on language use within radical communities has been carried out in the past by (van Heusden & Buis 1982). Although they did not have today's technology to guide them in their data analysis, they did find differences in language use such as repetition, noun use and belief stating in every publication (van Heusden & Buis 1982). Research by Hutton (1999) has focused on the degree in which a Nazi orthodoxy in terms of linguistics developed and on forums and how they communicate their ideology. Hutton (1999) showed that themes and ideologies of previous centuries are brought to life in fascistic texts. Research by Duffy (2003) looked into websites of four different hate groups. It has shown that some of the most prominent themes in extremist's texts fairness and morality of belief (Duffy 2003). Brindle (2009) used corpus techniques to analyse white supremacist language. Brindle (2009) found that 'homophobia, racism and sexism are inseparably interlinked' throughout white supremacist texts. Several other studies have looked into understanding motivations for radicalism through their content (Chertoff 2008, amongst others). A final study has been most successful to date in analysing language within radicalised groups by using quantitative techniques for text mining to find keywords and collocations in radicalised texts (Prentice et al. 2012).

Although the aforementioned studies yielded interesting results, they did not move past keywords analysis and did not take the growing body of online data available from social media into account. Until now, the majority failed to combine computational linguistics as well as new media methods to analyse their data. The aforementioned studies are small in nature and add work to a relatively scarce field of research. They, however, do show that linguistics features are salient and worth researching. Linguistic analysis provides a way to improve our understanding of radicalised language use. The current work builds on previous studies and tries to move beyond keyword analysis and focusses on a rapidly growing body of text that is popular amongst young users of the Internet. By doing so, it is expected to create a far more stable and up to date picture.

2.4 Machine learning

The study aims at implementing current state of the art technology in combination with computational linguistics to facilitate a classifier for text analysis. However, the implications of such a classifier are far wider than the scope of this thesis. Using a classifier to solve large text data problems for researchers falls under the category of machine learning. Computers are used herein to process data far bigger than can be done by hand. The goal of machines learning is to generalise patterns of analysis, or algorithms, for the future to create new knowledge out of data.

Algorithms have become a big part of our everyday life. They play in immense important role in what information is considered most relevant for users (Gillespie 2012). Algorithms map users' preferences and predict what they want to view. Work presented in this thesis only touches slightly on the notion of algorithms, but it is nevertheless important to review literature and take its implications into account.

2.4.1 Algorithms

Algorithms exist to make life easier, but many remain black-boxed to the outer world. An algorithm can therefore be best described as "any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values as output" (Cormen et al. 2009). Concisely put, an algorithm takes an input and transforms that to an output in a black-boxed manner. This is one of the reasons why algorithms have become a new concern for mundane people (Seaver 2014). For example, in order to analyse big chunks of data, a computer has to run an algorithm that makes predictions based on a given input. It is argued that they have the capacity to shape our ways of living and have a direct impact on individual lives (Barocas et al. 2013). Recent experiments have tried to unravel the mystery of algorithms, but have failed to date (Strathern 2000). Work within this field touches on sociological fields, wherein the consensus rests on examining more different algorithms to figure out their logic (Lazer et al 2009). One aspect that is mainly mentioned by Barocas et al. (2013), is that it is important to look at algorithms in a combinatorial way; no single algorithm is capable of producing all output, therefore studying algorithms has to be performed in a conjoined fashion within one working apparatus. All in all, algorithms are encoded procedures that transform input data into output data with meaning towards users, based on specified calculations. The current thesis has therefore chosen to work with an algorithm that can be mathematically explained and fine-tuned, which will be further elaborated on in the next section.

2.4.2 Document classification

One algorithmic application to process data is classification. This is built on a cognitive process that humans use to organise and apply our knowledge to the world (Cichosz 2015). A computer can run a classification model in which it has a representation of knowledge ready to project on new incoming instances, or data, to analyse according to the known set of attributes.

In terms document classification there is a distinction to be made. There are two main approaches for categorising documents: supervised and unsupervised learning. Supervised classification is based on building a model from a training set containing documents that have known categories. After the algorithm has build the model, it can predict the categories of input documents based on the known set of categories, or features. Contrastingly, unsupervised learning is mainly used in clustering where there is no need for a previous determined model. Herein, the algorithm uses a similarity method to cluster and categorise data (Rusu 2015). Section 3.4 will further discuss the chosen method of learning.

In order to make predictions about new data, an algorithm needs a set of features to determine its predictions on. There are multiple applicable methods available for classifying data. Options, amongst others, are a Bag Of Words (Ko 2012), wherein the number of words appearing across a document are counted, which results in vector build on word frequency hierarchy. Another option is the use of Support Vector Machines (SVM), (Pedregosa 2011). This uses vector representations of the entities in the training data. SVM splits the data using hyperplanes for classification (Statsoft.com 2016), which in its basis creates a multi-dimensional representation of the data. In other words, when sampling new data, it bases its prediction on the training data in terms of a linear classification. The method presented in this paper is the Naive Bayes method, which is build on Bayes' theorem and basis itself on strong independence assumptions between feature (Rusu 2015). Fortunately, it is less black-boxed than other methods since it can be explained with one equation. Because this method does not have a complicated iterative process, it is perfectly suitable for running large datasets as will be shown in section 3.4.

3. Approach

Technological innovation has dramatically increased the type and range of documents open to qualitative analysis. As Rogers (2013) points out in digital methods, there should be more focus on interaction patterns as part of virtual communities in social networking. He also states that when researching new media platforms one must pay attention to the social elements such as language or communities (Rogers 2013). Analysing social media can result in obtaining important information on the community and its events. Users within certain platforms can share this sense of community via their posts, resulting in an online social well-being (Zappavigna 2011). As

aforementioned, O'hara & stevens (2015) describe this as the "we-sense". In order to reach this, researchers need to adapt a new and different form of research in which methods are suitable and applicable to the new paradigm. This can be achieved by combining community theories with the aforementioned linguistic phenomena. In light or Rogers (2013), data are not just opportunity datasets, but pose a chance to conduct original and meaningful research. The current work builds on Roger's (2013) notion and plea for new research methods in which they are applicable and sustainably usable for online social research.

Internet sources, such as the discourses created within virtual communities, are especially open to Roger's (2013) approach whereby they can be systematically examined and assessed to develop an understanding of their complex social phenomena. The endeavour in this work is to create a proof of concept for classifying radicalised text within the online Reddit network. Reddit is an entertainment website where registered members can submit, comment or like ("upvote") content in forum-like style. The entries of content are divided into so-called sub-reddits, which include various topics as gaming, news, movies, books and photography amongst many more. The used corpus consists of known radicalised and non-radicalised sub-reddits. Linguistic corpus analytics is chosen as the representative analysis in combination with machine learning, since they will provide occurrences that count as factual evidence of language taking place (Biber et al. 1998). This results in a computer automated classification algorithm.

The approach proceeds in four steps. The first step, as described in section 3.1, involves sampling the data to equal out the two Reddit corpora. Secondly, both texts will be cleaned and pre-processed, as shown in section 3.2. Thirdly, section 3.3 describes the general statistics and text mining will be executed for analysis by means of part-of-speech tagging, frequency distribution and collocation (bi-gram) analysis. Lastly, the texts will be implemented into a machine learning task TD-IDF based on Naive Bayes for document classification as described in section 3.4 (Rajalakshmi & Aravindan 2011).

3.1 Data acquisition

Data was gathered as part of the VOX-Pol[3] project and performed at TNO The Hague[4]. The original dataset is crawled via the Reddit API from October 2007 to May 2015, containing 105,930,239 topics divided over 239,773 sub-reddits. The dataset contained clear noticeable radicalised sub-reddits (van Gysel & de Rijke 2015). An overview of the sub-reddits contained in the radical file can be found in table 1. The sub-reddits are diverse in topic, which make this dataset perfectly suitable for the current work.

[3] A European Union Programme for academic research focusing on violent political extremism.
[4] Netherlands organisation for applied scientific research.

Table 1. Overview of Sub-Reddits Containing Controversial Content		
Sub-reddit	Description	Posts
/r/AntiPOZi	Pro-white community	28,238
/r/ShitRedditSays	A sub-reddit which often criticizes controversial content on Reddit	967,311
/r/SRSsucks	Anti-Shit Reddit Says	311,143
/r/antisrs	Anti-Shit Reddit Says	151,680
/r/new_right	Far-right extremism	20,673
/r/FeministHate	Feminist hate	9
/r/Polistan	Anti-semitism	3,485
/r/GasTheKikes	Anti-semitism	1,887
/r/WhiteRights	White extremism	78,004
/r/nationalism	Nationalism	229
/r/WhiteNationalism	White nationalism	854
/r/farright	Extreme-right	100
/r/CoonTown	Racism (*banned as of August 2015*)	199,176
/r/Anti_Gay	Anti-homosexual sub-reddit	82

Table 1. Overview of Sub-Reddits[5]

The non-radical file originally contained 53.851.542 lines and the radical file contained 1.7 million lines. Both files were sampled down to 1 millions lines to limit processing difficulties for the system. Cut-offs for the sample were set to a minimum of 1500 for keywords and 200 for collocations; this means that words presented in the analysis appear at least 1500 or 200 times per given category. The files were sampled randomly to match layout criteria, such as chronological order. A second sample was drawn consisting of 50.000 lines per dataset which was only used to create dispersion plots[6]; these plots visualise occurrences of words over time. This was chosen for, since 50.000 is the maximum number of lines that can be processed without losing overview in the plot. Not cut-offs in terms a minimum of occurring keywords or collocations were used.

3.2 Data pre-processing

Corpus linguistics involves analysis performed on large documents of text. Before analysing the bodies of text in terms of keywords or frequency distribution, the text was cleaned on redundant information. The posts from the radical and non radical files were loaded into Python[7] and read as strings of text. Cleaning was performed with the NLTK[8] and SciKit[9] library. NLTK stands for *Natural*

[5] van Gysel and de Rijke 2015
[6] As described in section 3.3
[7] https://www.python.org
[8] http://www.nltk.org
[9] http://scikit-learn.org/stable/

Language Toolkit and is one of the most used library for natural language processing. Functions included in NLTK and SciKit can transform all words to lower case and delete stop words for example. All words were changed to lower case to improve accuracy of the analysis and to successfully perform stop word deletion. A stop word filter discards the words of little or no relevance within the dataset; this includes words that are frequently used in a language as “the” or “a” in English. Additionally, words of with a length of 1 or 2 characters were deleted from the corpus since these have less meaning in bigger texts. A custom stop word list was provided since the SciKit stop word list only holds 318 stop words. The custom stop word list is a combination of NLTK, SciKit stop words and a list by Buckley and Salton (2016) with a total of 871 stop words. A custom list as presented here results in a more thorough deletion of redundant words in the text, which improves the analysis as well as the performance of the script[10]. A Porter Stemmer[11] is used to reduce all words to their root or stem. This improves accuracy on retrieving word count and collocations from texts since less mistakes between nouns or verbs are made. For example, a stemmer changes words as “fuck” and “fucking” to “fuck” which improves accuracy for the purpose of the current thesis. Also, it improves the accuracy of the part-of-speech tagger in displaying a more truthful picture on the amount of verbs and nouns within the datasets. Finally, the script uses regular expressions, a pattern matching library, to find and delete URL’s and email addresses that are redundant for the current thesis. The entire code can be found in the appendix under section 9.

3.3 General statistics and text mining

The clean radical and non-radical datasets were further analysed on different language features to get a clearer picture of the language within both corpora. Firstly, keywords were counted and stored into a frequency distribution. As aforementioned, words occurring more than 1500 times in the text were taken into consideration. The top 1000 words of the frequency distribution was then written to an external file to provide an overview of the most common words in both files. The number of most common words was set to 1500 as this provides an overview of most used words from both texts. For the classifier, the amount of occurring words was set to 10, since lesser occurring words play an important role for the classifier as explained in detail in section 3.4. Additionally, a file with the frequency of words with a length longer than 7 characters was produced as this provides insights in the division of long words per text. Choosing too many would make it too hard to analyse manually, while too little would result in displaying common words that were not deleted by pre-processing, but still hold little linguistic value for this thesis. Subsequently, the most common words (collocations) were plotted in a dispersion plot and stored in an external file based on the 50.000 lines sample. A dispersion plot reveals (co)-occurrences of words over the

[10] The complete list of stop words can be found at https://goo.gl/SigTUN
[11] http://www.nltk.org/_modules/nltk/stem/porter.html

entire document. Lexical diversity was obtained over all occurring words to signal the diversity in word choice per text. Words were tokenised using the NLTK tokenise function and subsequently fed to a part-of-speech tagger. The NLTK part-of-speech tagger scans all the sentences and words in the texts and assigns labels to them, as shown in figure 1.

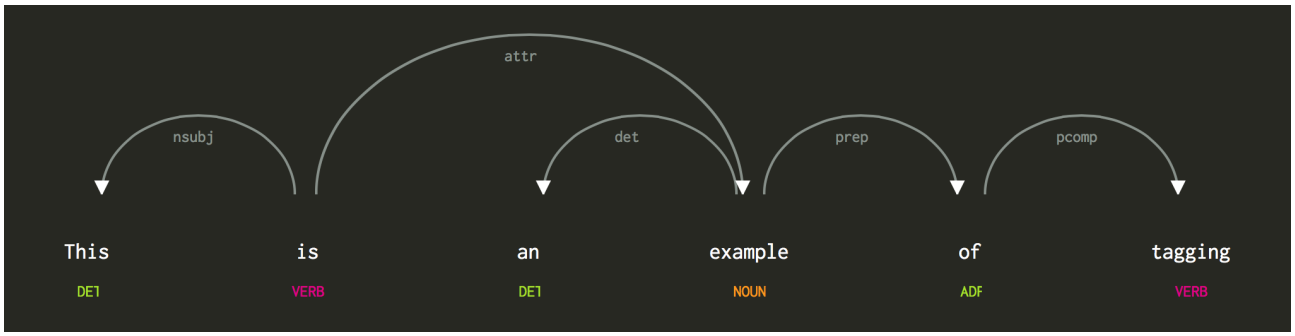


Figure 1. Example of NLTK part-of-speech tagging

The current analysis performed with the part-of-speech focussed on the use of nouns, verbs and pronouns per dataset. The occurrences were stored in an external file for analysis. All the selections above have been made in relation to the the main focus of this thesis, which is on document classification. The numbers and outcome files mentioned above are merely used to inform us with insights from the texts themselves based on the aforementioned selected features.

3.4 Corpus analytics

The loaded texts were labelled as radical or non-radical in language nature and stored in a feature set with radical and non-radical labels. The feature set was subsequently loaded into a machine learning algorithm based on TF-IDF to train a Naive Bayes Classifier. Figure 2 schematically outlines the machine learning algorithm.

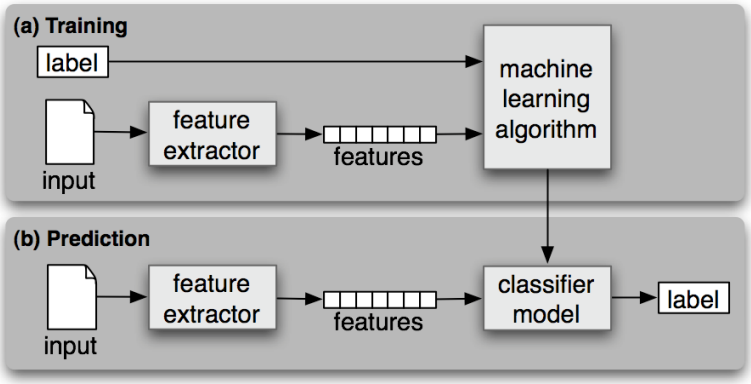


Figure 2. Schematic machine learning overview[12]

[12] From: http://www.nltk.org/book/ch6

The Naive Bayes classifier takes a training and a test (prediction) set into its algorithm. The division chosen for this thesis is 40/60, which means that 40% of the documents is used to train the algorithm and 60% to test the trained algorithm. The division is performed randomly on a combination of both texts (radical / non-radical). The features are extracted and stored in combination to the occurrences into the learning algorithm (Figure 2, (a)). The test phase does a similar job but does not come with labels, merely its features are extracted (Figure 2, (b)). Finally, the model returns a set of labels as outcome for the test set as well as an accuracy score of labelling correctly.

A Naive Bayes Classifier is based on the so-called Bayesian theorem and is chosen based on its performance with high dimensional input. TF-IDF stands for *term frequency-inverse document frequency*[13]. It places all the words and its labels taken from the corpora in a vector and takes placement and occurrence into account. It does so by assigning weight to the occurrences, which provide a more reliable basis for analysis and classifying that boosts less occurring words instead of just looking at frequencies. The classifier takes a text or line of text as an input and mirrors that the known datasets to check similarities of that text. Subsequently, the classifier checks the labels and predicts whether or not the text is radical or non-radical in terms of language use. The formula for the Naive Bayes classifier can be noted as:

$$g(x,l) = \begin{cases} 1 & \text{if } f_i = a, (i = 1, \ldots, k) \\ 0 & \text{otherwise} \end{cases}$$

The formula takes an input g, that contains of x (the text) and l (labels). If they match, it returns 1 wherein the majority of occurrences in the text match the labels and connected words from the known texts. If not, it returns 0. In other words, it takes a target (texts that are assumed to be radical in nature) and returns a prediction based on previous known features, or known texts.

The algorithm basis is calculations on supervised learning, which for the current work is the labelled set of texts from whereon it predicts the outcome. Additionally, a standard Naive Bayes classifier only counts words within sentences and does not assign weight to the occurring words and their position. It was therefore chosen to implement the TF-IDF application to the Naive Bayes algorithm. The weight is a statistical measure that evaluates the importance of a word to a document in a corpus (Buckly & Salton 1988). The importance of a word increases proportionally to the number of times a word appears in the document. It is then offset by the frequency of that words in the corpus. In other words, it ranks the words in the corpus according to its relevance within the document and indicates this by means of number. Moreover, rare terms are boosted since they are weighed. This creates a more evenly divided picture of the corpora. The end result

[13] http://www.tfidf.com

is an accuracy, the fraction of correct predictions of the classifier in percentages, and a list of most informative features on which the algorithm based its choice on (Buckly & Salton 1988). The aim of the classifier is to classify in terms of precession and less on recall (Davis & Goadrich 2006). This matches the goals of the thesis in terms of a proof of concept. A focus on precision entails finding documents that contain radicalised texts. A focus on recall would entail finding all the radicalised texts, which can lead to false positives (David & Goadrich 2006). Precision was chosen as a goal for this thesis; the classification will focus on correctly assigning radical content a radical label, since the evidence of finding and recognising radicalised discourse is key for the current work. The amount of most informative features has been set to 1000 features as this will provide a solid overview of features that the algorithm based its prediction on.

The texts used were selected with care, but only contain topic including with the Reddit platform and no information from other social media platforms. This was chosen for since this provides a more reliable picture for a proof of concept when testing with similar data and building a corpus of radicalised language online.

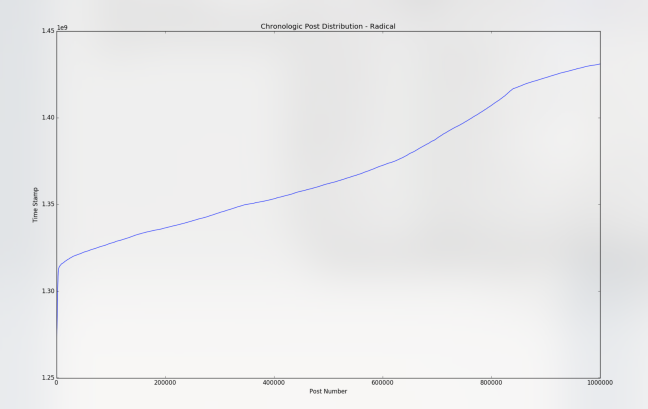
4. Results

The following sections provide an overview of the obtained results from this study. Section 4.1 provides descriptive statistics containing chronological order of the samples, word count and lexical diversity. Section 4.2 lists results from the word frequency. Section 4.3 displays the collocations and dispersion results. Section 4.4 builds on results obtained from the part-of-speech tagger. The final section, 4.5, provides results obtained from the document classifier as accuracy and most salient features[14].

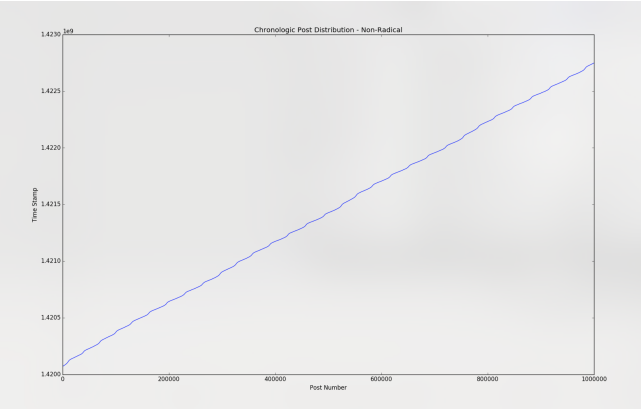
4.1 Overview of datasets

As described in the method section, the posts are randomly sampled from the datasets. Plot 1 and 2 on the next page prove that the sampled data is represented in chronological order. The horizontal axis displays the amount of lines and the vertical axis displays the time stamp. This is an important aspect for further analysis since the discourse is still in the same order as the original. Data in both plots show a stable linear line across both corpora.

[14] The complete list of tables is available at https://goo.gl/SigTUN



Plot 1. Chronological radical distribution



Plot 2. Chronological non-radical distribution

Table 2 presents an overview of the data contained in the sample set of 1 million lines. The amount of unique words does barely show a difference. This means that the use of words is equally unique in the radical and non-radical dataset. Interestingly, the data show a large difference in the amount of collocations between radical and non-radical. This indicates that the radicalised discourse needs more collocations to convey their message than non-radical discourse. It shows that radicalised communities on Reddit use more collocations than non-radical communities. Section 4.3 will provide deeper analysis on this topic. The table also displays that there is little to no difference in lexical diversity and thus not significant on its own. This shows that both groups are similar in their amount of unique words. Finally, the data show that the word count is significantly higher in the radical dataset compared to the non-radical dataset. This, in line with the collocations, indicates that radicalised communities use more words to convey their message than non-radical communities.

Overview Data – 1 million lines					
	Label	*Unique Words*	*Collocation Count*	*Lexical Diversity*	*Word Count*
	radical	1.130	1.071	0,000171964	6.571.130
	non-Radical	1.063	419	0,000210528	5.049.217
Difference		67	652	-0,000038564	1.521.913

Table 2. Overview of the radical and non-radical datasets

4.2 Word frequencies

By comparing the radical corpus with the non-radical corpus, it was possible to establish similarities and differences across the results. These results provide an indication of the key words that are specific for radicalised texts. In other words, if a particular word appears more in the

radicalised corpus this could be indicative for features of radicalised language use. Table 3 shows the top 35 results of keyword distribution for radical posts and non-radical posts.

Radical	
count	*word*
83680	fuck
78095	women
77028	white
71063	reddit
63813	sr
57856	time
52324	black
50876	shit
48952	person
46856	rape
46172	guy
39130	feel
36827	hate
31152	woman
29578	racist
29094	pretti
29056	thread
29024	talk
28821	nigger
28246	joke
27917	histori
27821	word
27657	sex
27559	read
27472	bad
26295	love
24761	real
24192	day
23851	live
23790	girl

Non-radical	
count	*word*
76032	time
57739	game
43893	play
34011	feel
33001	day
32736	fuck
31634	pretti
30988	love
29242	start
29237	guy
25918	person
21818	shit
21634	bad
20875	question
20753	read
20649	live
19521	talk
19317	friend
18604	team
18399	watch
17672	life
17513	bit
16343	sound
16176	money
16019	idea
15966	player
15871	hard
15759	nice
15216	buy
14814	kill

23012	cultur
22917	world
22903	wrong
22751	feminist
22601	understand

14687	set
14548	world
14436	week
14420	edit
14183	stuff

Table 3. Overview top 35 most common keywords

As table 3 shows, keywords in the radical dataset score higher compared to non-radical keywords; this indicates that the radical dataset contains more similar word use within its community. These results are in line with word count shown in table 2. What is interesting about these results is the amount of difference in keywords and count of keywords between the two datasets. For example, the words used in the radical dataset as “white”, “rape”, “hate” and “nigger” provide a vivid picture into keywords used within the radical community. Also, these words only occur in the radical dataset. It indicates that their word use is significantly different in terms of keywords compared to the non-radical dataset. This is likely due to the different nature of topics for radicalised discourse. The data shows that the radical dataset therefore contains salient features typical for their discourse in terms of keywords.

4.3 Collocations and dispersion

Table 4 provides an overview of the top 35 collocations in the dataset. It summaries collocations for the one million lines samples.

Radical	
word	count
vote histori	13060
screenshot vote	13045
free speech	3913
social justic	3229
white male	3174
holi shit	3065
video game	2865
histori histori	2814
real life	2726
rape cultur	2714
fuck fuck	2453

Non-radical	
word	count
action perform	4477
perform automat	4463
contact moder	4406
automat contact	4379
play game	2987
game play	1763
answer question	1732
feel free	1613
holi shit	1527
amp amp	1449
time time	1400

feel bad	2082	video game	1386
sexual assault	2075	spend time	1276
black white	2073	month ago	1261
rape joke	2018	submiss remov	1048
rape victim	2006	feel bad	1023
white guy	1984	real life	1010
white person	1873	wast time	1006
piec shit	1867	automat remov	955
hate women	1858	week ago	912
snapshot readabl	1830	hard time	903
gender role	1784	game game	850
black person	1755	day day	827
child support	1733	pretti cool	825
spend time	1694	pay attent	821
nigger nigger	1689	day ago	808
fals rape	1682	read book	783
fals accus	1645	spend money	756
polit correct	1602	god damn	747
affirm action	1571	submiss automat	742
real world	1571	origin submiss	735
child porn	1528	time play	708
white white	1517	time day	706
straight white	1513	volunt tribut	698
downvot brigad	1487	gather mc	690

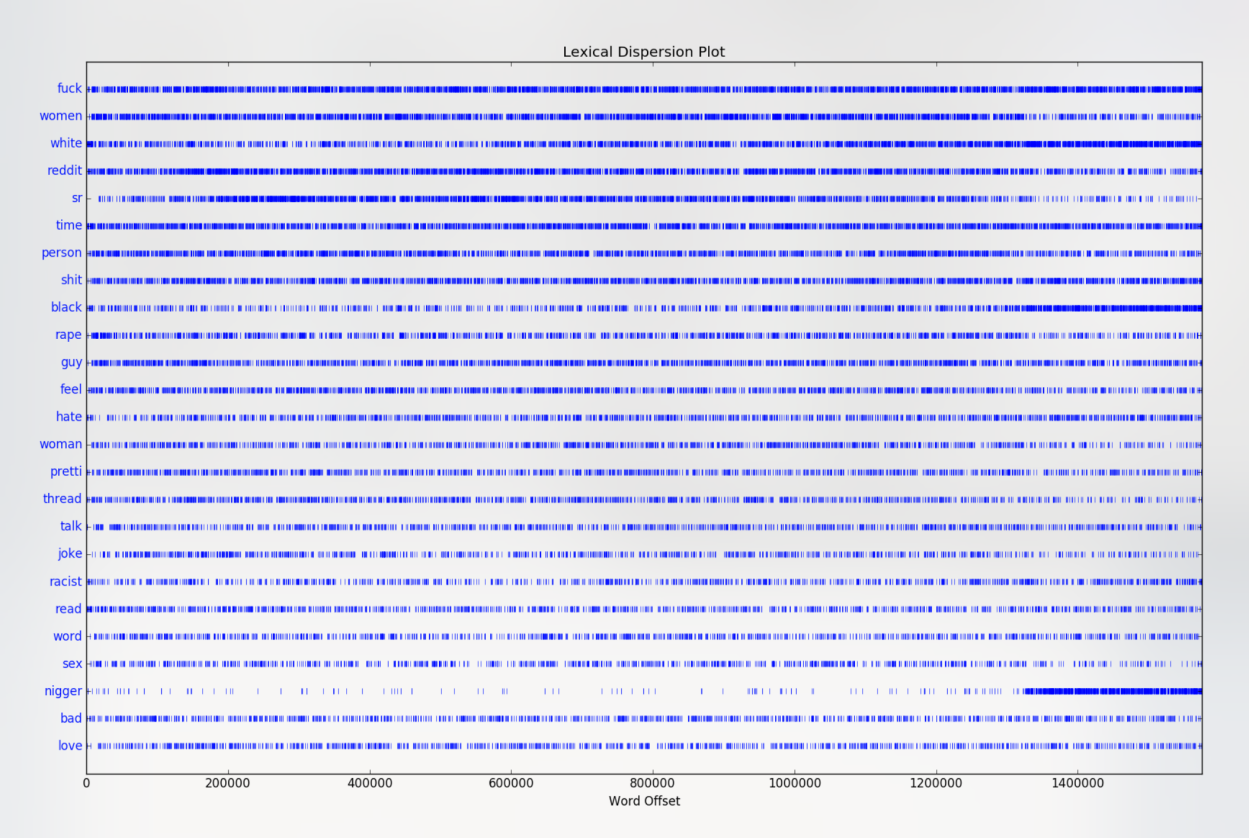
Table 4. Overview top 35 most common collocations

Interestingly, table 4 shows that overall collocations appear more often in the radical dataset than the non-radical. This has been shown in table 2 in terms of collocations, but is also visible in table 4 above. The amount of collocations for the radical discourse do not drop below 1645 occurrences in the top 35, while in the non-radical dataset it drops below 800. Also, the top 35 collocations displayed shows regularly used, and vivid, language use within the community. The amount of collocations shows like-minded language use in terms of using words together (Pennebaker 2011). The type of word combinations used is completely different between both communities which provides insights into the radical discourse in terms of collocations. For example, words as “white male”, “sexual assault”, “black person” and “child porn” are frequently used in combination throughout the radical dataset. When looking at the number of occurrences the data show that collocations as “hate women” appears 1858 in the radical dataset. This means that “hate women”

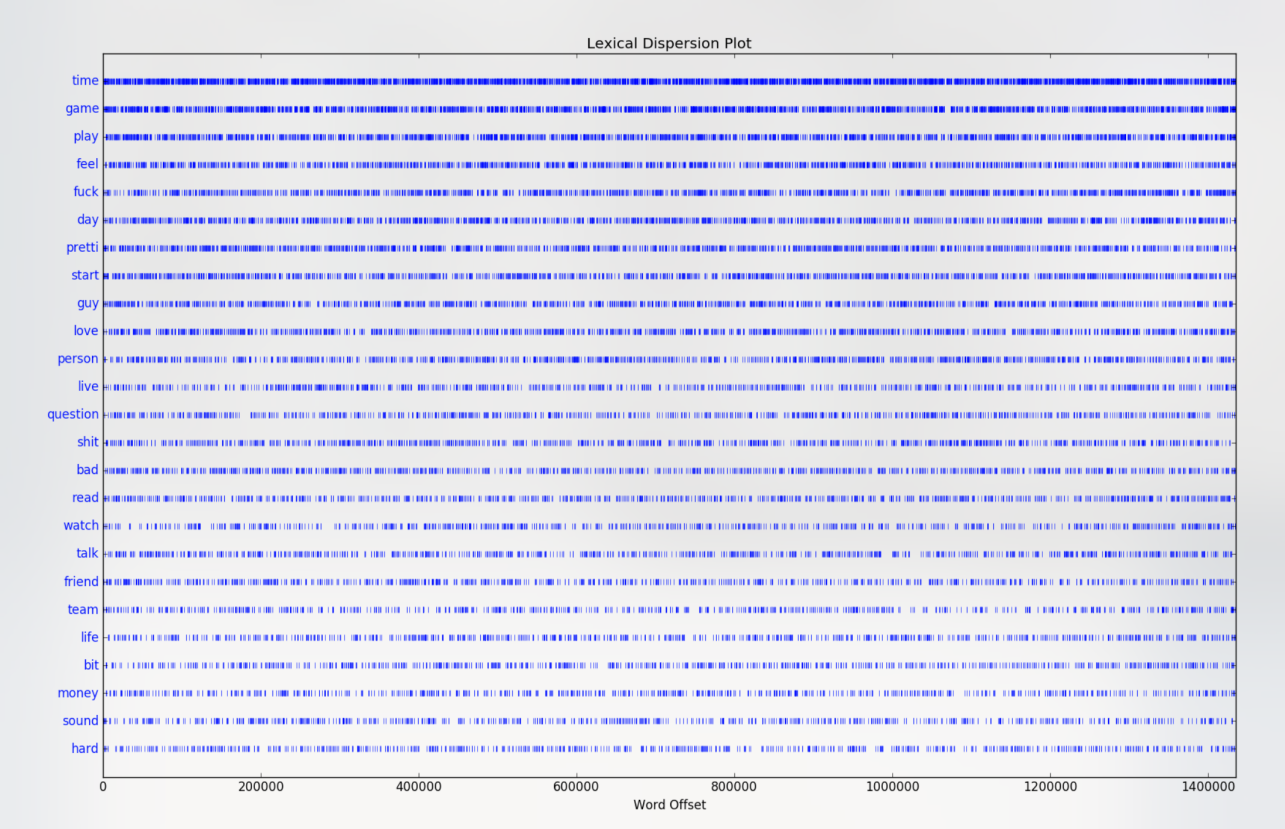
appears in 0,18% of the total radical dataset. Another interesting example is “free speech” which appears 3913 times in the radical dataset, roughly taking up 0,4% of the dataset. On the other hand, the non-radical dataset contains more mundane words used for collocations as “play game”, “pretty cool” and “save money”. Therefore, the collocations in the radical dataset contain important features typical of their discourse within their community. The classifier builds on these features, which is explained in section 4.5.

4.3.1 Dispersion plots

The dispersion plots displayed in plots 3 and 4 show words used over time. These plots have been generated from the 50.000 lines sample and provide an interesting window into the occurrences of the top 25 most frequent keywords over time, between October 2007 and May 2015. The amount of blue bars and the thickness signal the occurrence of words in the discourse.



Plot 3. Lexical dispersion plot radical discourse



Plot 4. Lexical dispersion plot non-radical discourse

As plot 3 shows, the word “nigger” did not appear frequently until the end of the dataset. This means that words used within the communities not only change over time, but are also heavily accommodated to by other members in the radical community. Taking into account the late occurrence of the word “nigger” in the dataset, it appears 28821 times, or in 2.88% of the radical dataset. Furthermore, the plot shows collocations, or co-occurrences over time. As shown, the word “nigger” appears together with words as “white” and “black” at the end of the dataset more often. This proves not only a difference in keyword use over time, but also collocations being used more frequently within the discourse of the community. The word “fuck” also seems to be an indicative keyword for the radical discourse, as it appears on a steady level throughout the discourse and in many collocations. On the other hand, the dispersion plot the non-radical dataset shows a widespread division without any clear patterns of frequently used keywords over time within their discourse. Moreover, it displays a stark difference in collocations since no patterns can be found of frequently words used together frequently compared to the radical discourse. The radical discourse displays more like-minded language use within its community. This, again, shows that the radical dataset contains salient features of language within its discourse.

4.3.2 Long keywords

Table 5 shows an overview of the top 25 keywords longer than 7 letters. The data show a similar pattern in the use of longer words, compared to the overall keyword use as shown in table 3.

Radical	
count	*word*
27917	histori
22751	feminist
22601	understand
20794	redditor
19054	subreddit
17602	screenshot
15988	societi
15724	internet
15271	discuss
15236	support
14988	argument
14468	children
14466	privileg
14429	question
14385	complet
14127	opinion
14010	downvot
13835	respons
13150	countri
12537	oppress
11587	shitlord
10974	attract
10817	american
10514	bullshit
9886	account

Non-radical	
count	*word*
20875	question
13873	understand
12176	complet
10636	support
9044	charact
8648	perform
8636	control
8398	countri
8235	mention
8206	contact
7731	account
7391	compani
7369	absolut
7299	opinion
7210	automat
6511	respons
6273	product
6231	honestli
6229	explain
6065	version
5997	favorit
5975	discuss
5886	recommend
5854	correct
5796	develop

Table 5: Overview of top 25 most common long keywords

Despite the results showing a less clear picture of specific radical language use in the case of longer words, the results still add up to the results shown in table 3 and plot 3 and 4. Words as “privilege”, “oppress” and “feminist” still provide a clear picture in terms of keywords used compared to the less interconnected set of longer keywords displayed within the non-radical discourse. It shows that case of short and longer keywords their language use accommodates in terms of count and. The data portray a stable picture of stark differences between the groups in language discourse as well as showing language accommodation within the radical community.

4.4 Part-of-speech tagging

The NLTK part-of-speech tagger moves beyond keywords and analyses the amount of word classes used in both datasets. Table 6 provides an overview of the amount of nouns and verbs used in both datasets.

Label	**Nouns**	**Verbs**
Radical	4259562	527576
Non-radical	3319793	526791
Difference	28%	0,04%

Table 6: Overview part-of-speech tagger results

Table 6 shows that there is a stark difference in the amount of nouns; there are 28% more nouns used in the radical dataset than the non-radical. The difference in the use of verbs is negligible. The difference in noun use displayed above builds upon earlier results shown in the keyword analysis. It shows that the radical discourse uses more nouns to convey their messages than in the non-radical dataset. These results overlap with the outcomes of the collocation analysis as shown in table 4. It seems that the radicalised community needs more nouns to convey their messages online, which is connected to collocation use, as these mainly consist of nouns. All in all, the part-of-speech tagging data accounts towards the word count as shown in table 6 and the overall use of keywords as shown in table 3. The data presented here show difference in the amount of nouns, the use of keywords and provide salient features of language difference used within the discourse of the radical community.

4.5 Document classifier

The document classifier is the final stage of the analysis for the radical and non-radical dataset. The results shown in previous sections provide an interesting overview of the data contain in the datasets. The document classifier builds on the previously displayed results and takes salient language features into calculation towards the probability of a text belonging to the radical or non-

radical discourse. As explained in the method, the classifier takes the content of both datasets and labels to assign labels to new incoming data. The classifier was able to perform with 75% accuracy on correctly labelling documents as belonging to the radical or non-radical discourse. In terms of accuracy, it performed above chance level. In case of a non-working classifier, the accuracy would result in 50%, thus based on chance. The current classifier was able to correctly label radicalised discourse with a focus on precision. Table 7 displays the top 25 most informative features as chosen by the classifier.

Label	Word	Probability
Radical	fuck	-5,031980011
Radical	reddit	-5,115399634
Radical	white	-5,252443166
Radical	sr	-5,255207419
Radical	women	-5,289180854
Radical	shit	-5,494117237
Radical	time	-5,630065005
Radical	rape	-5,641316732
Radical	guy	-5,652046021
Radical	black	-5,663988477
Radical	hate	-5,779670562
Radical	person	-5,78910785
Radical	feel	-5,852145589
Radical	thread	-5,852752395
Radical	nigger	-5,861524
Radical	joke	-5,865014532
Radical	racist	-5,886168287
Radical	lol	-5,952195194
Radical	read	-6,008968364
Radical	love	-6,028734104
Radical	woman	-6,030201921
Radical	pretti	-6,04838274
Radical	word	-6,064271697
Radical	redditor	-6,10111855
Radical	bad	-6,105652194

Label	Word	Probability
Non-Radical	abathur	-18,70850662
Non-Radical	abi	-18,70850662
Non-Radical	ableton	-18,70850662
Non-Radical	abra	-18,70850662
Non-Radical	abzan	-18,70850662
Non-Radical	acasi	-18,70850662
Non-Radical	acceleromet	-18,70850662
Non-Radical	acho	-18,70850662
Non-Radical	actuat	-18,70850662
Non-Radical	acuerdo	-18,70850662
Non-Radical	acum	-18,70850662
Non-Radical	acx	-18,70850662
Non-Radical	adaptor	-18,70850662
Non-Radical	adb	-18,70850662
Non-Radical	addedsuperimpos	-18,70850662
Non-Radical	adelant	-18,70850662
Non-Radical	adem	-18,70850662
Non-Radical	aeg	-18,70850662
Non-Radical	aegislash	-18,70850662
Non-Radical	aero	-18,70850662
Non-Radical	aerotank	-18,70850662
Non-Radical	afccg	-18,70850662
Non-Radical	afl	-18,70850662
Non-Radical	aftermarket	-18,70850662
Non-Radical	aftershav	-18,70850662

Table 7: Overview of top 25 most informative features

Table 7 displays the most informative features on which the classifier based its classification. It shows a similar distribution as laid out in the keyword analysis: the non-radical discourse shows more mundane words compared to the radical discourse. It shows that the keyword “fuck” is roughly 5 times more likely to appear in a radical discourse than the non-radical. Also, words that became apparent in section 4.3 in the dispersion plots also prove informative for the classifier. Words as “rape”, “black” and “hate” all have a higher probability of appearing in the radical discourse than the non-radical discourse. On the other hand, words as “adaptor” or “aftermarket” are amongst the highest probability numbers for labelling a text as non-radical. The most informative words from the non-radical dataset do not provide a general picture of features

contained in the dataset, whereas features from the radical dataset provide a strong image. These data show that the classifier operates on an acceptable accuracy level, but also show that the radical discourse contains salient language features that can be used for classifier purposes.

5. Discussion

The rise of social media platforms and the dynamic nature of language provide an interesting and promising source of information for research. Within these platforms, language is objective and quantifiable data that allows researchers to study online behaviour as users present themselves in their natural, unique way (Hauffa et al. 2014). A combination between more traditional and new multidisciplinary methods can bridge the gap to making sense of the big pools of data that are available online (Rogers 2013). The current approach offered an immersive proof of concept method for recognising online radicalised language within the Reddit community. This thesis has taken Reddit as a research corpus to perform keyword, collocation, part-of-speech tagging and machine learning techniques to answer the following question: which distinctive patterns of language can be found for language coordination within radicalised online forum communities? In order to answer the main question, this paper addresses the following sub questions: What is the frequency distribution in the dataset? What are the most common keywords within the dataset? What are co-occurrence patterns for words? How do they compare to other posts within non-radicalised sections of the forum?

Returning to the main research question posed in this thesis, the data show that the keyword frequencies differ between the radical and non-radical communities. The radical community reveals more similar keyword usage as than the non-radical community. This like-minded language use reveals not only the frequency of their keywords, but also the kind of keywords they use. In light of Pennebaker (2011), the keywords used in the radical community provide a window into their language use as a virtual community. The use of collocations shows a similar distribution; the radical community uses more collocations than the non-radical community. In combination with the dispersion analysis it shows that the language use within the radical community is heavily influenced by its participants. The results confirm that social stratifications play an important role not in the use of Internet language, but also in the perception of language (Crystal 2001). The linguistic behaviour as part of groups is an indicator of social presence within a virtual community.

In line with Bowman-Grieve's (2009) research, the current work shows that active participation is key in forming an online community. Results in this work have shown that the radical community builds on participation and can be viewed in light of Rheingolds' (2000) virtual community theorem. The results have shown that the radical community is not only active as a

community, but also converging in their way of language use. Their language use accommodates towards the most prevalent language use in the community. The data show that their language use in terms of register accommodates towards the general topics within the group, as shown by the rise of the word “nigger” and “black” near the end of the dataset. The language use within the radical community shows that they use it to portray their standards and ideologies (Gruzd & Roy 2014). They display diversity from the non-radical communities in terms of language use; thus, creating their own echo chamber (O’Hara & Stevens 2015; Rogers 2004). The data also show that virtual communities act in a similar way; the notion of community is less apparent in the non-radical discourse, which is shown by the difference in language patterns compared to the radical discourse (Rheingold 2000). The online platform shows its added value for sharing shared values, as the results show for the radical community (Passman et al. 2014).

The new media and corpus linguistic techniques demonstrated in this thesis show only a fraction of the possible outcomes of combining these fields. Other methods, such as combining network analysis to find authors to match text to, stylistic analysis (matching writing to authors), topic modelling (revealing general topics within texts) or more thorough dispersion analysis could be incorporate to improve the existing body of knowledge. This could prove useful for further analysis of radicalised texts and is an interesting field for future work. Furthermore, adding more and different data sources would not only move the current work beyond a proof of concept, but could also lead to group comparisons and training the classifier more thoroughly on different sources of data. This could ultimately lead to a predictive algorithm as part of a crawler to independently search the web for radicalised content.

With the future implications aside, it should be noted that the current work is not without its limits. Firstly, corpus analysis tools are still in their developmental phase. Secondly, the data pool used is only a small portion of the available data online. Furthermore, this work has presented a proof of concept which in the analysis lead to a bird’s eye view of the data, leaving interpretation mainly in the hands of the researcher; this may result in a researcher bias. Finally, other classifier methods could be tested and pre-processing of the texts could be addressed to improve overall performance of the classifier (Statsof.com 2016). Fortunately, the techniques described in this work are replicable for other researchers to reach similar conclusions.

Our understanding of the language use of radical virtual communites has increased with this work. In light of previous research that focussed on keywords and older sources, this work contributes and continuous work in an important field of study (Burris, Smith & Strahm 2003; Gustavson & Sherkat 2004; Attestog & Perera 2013). It furthers the use of online sources and looks past traditional fields of study by providing a multidisciplinary study with a focus on an up-to-date source. This work has shown that salient language features can be discerned from large amounts of text. It has also shown that language use differs in the radical and non-radical discourses. This has lead to a classifier correctly predicting a community’s discourse as radical or

non-radical in 75% of the cases. Since the aim of this research is precision, it means that the classifier in this thesis performs well above chance. It is therefore capable of detecting radicalised discourse within the Reddit platform. An important note must be added; the classifier only labels a text as radical when it is evidently radical. This means that not all radical texts will be recognised. In other words, the current classifier functions as a detection tool, whereas future research can focus on improving the recall of the classifier. Improving recall will result in a higher retrieval of radical documents, but can also lead false positives (Davis & Goadrich 2006). Future research thus has to careful take the trade-offs into account when improving the accuracy of such a classifier.

The language used within the communities is an important aspect of virtual community building. This not only shows a 'we' sense as part of the sub-reddit, but also displays this in terms of language use within the radical virtual community (Anderson 1991). The data have shown that the language use within the radical communities accommodates over time, resulting in higher scores of keywords and collocations. Also, as many keywords are taken as most informative features for the classifier used, it shows that their use of language within the communities is a sign of their community. This results in similar language use, making it possible to abstract their salient language features.

6. Conclusion

The techniques demonstrated in this thesis show only a fraction of the possible outcomes of corpus linguistics. This work has performed an analysis by means of combining the fields of new media and linguistics. It has done so by performing a proof of concept for classifying online radicalised from non-radicalised texts. The implications of such a classifier are far wider than the scope of thesis; a proof of concept can be taken as a basis for an algorithm as part of a bigger whole wherein online radicalisation can be tracked or monitored.

In light of Rogers (2013), this work shows that multidisciplinary methods yield interesting results and should be further explored. The work presented in this thesis has succeeded in building on previous theories from the fields of new media and linguistics by adapting methods from both domains for analysing online community behaviour in terms of language. Their use of language has resulted in salient language features, which could be used by the presented machine learning algorithm to predict and differentiate radical texts from non-radical texts.

The combination of keyword frequency, collocation analysis and part-of-speech tagging has provided this work with a window into the discourse of radicalised virtual communities. Combined with the classifier, it has allowed the current study to take a varied approach to the data. It has provided the researcher with consistent observations wherein the analysis as well as the classifier confirmed the presence of salient language features within the discourse of radicalised virtual

communities on the social media platform Reddit. This opens the possibility for a scalable technique to not only improve research into radicalised communities, but also to move past them and analyse other groups of interest. The techniques laid out in this thesis are therefore useable for much larger datasets. This thesis may also inform future analyses for caveats and improvements on the current method. Summarising, the thesis highlights the value gained from combining theories and techniques from the fields of new media and linguistics.

7. Acknowledgements

This project has been supported as part of the VOX-Pol project at TNO The Hague. I would like to thank Mirko Tobias Schäfer and Stefan Werning from Utrecht University and Gijs Koot and Maya Sappelli for their insights, feedback and support regarding this work.

8. Literature

Anderson, B. 1991. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London, UK: Verso.

Anger, I., & Kittl, C. 2011. "Measuring influence on Twitter". *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies* - I-KNOW '11, 1. doi:10.1145/2024288.2024326

Attestog, T., & Perera, S. 2013. "Mapping Extremist Forums using Text Mining". AURA.

Aw, A., Zhang, M., Xiao, J. & Su, J. 2006. "A phrase-based statistical model for SMS text normalization." *In Proceedings of ACL,* pages 33–40.

Balcerzak, B., & Jaworski, W. 2015. "Application of Linguistic Cues in the Analysis of Language", *16*(2), 145–156.

Barocas, S., Hood, S., Malte, Z. 2013. *Governing Algorithms: A Provocation Piece.*

Barton, D., Lee, C. 2013. "Language Online: Investigating Digital Texts And Practices". *Studies in Second Language Acquisition*, 224. http://doi.org/10.1017/S0272263114000254

von Behr, I., Reding, A., Edwards, C., & Gribbon, L. 2013. "Radicalisation on the digital era. The use of the internet in 15 cases of terrorism and extremism". *RAND Europe*.

Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University press

Biber, D., & Conrad, S. 2009. *Register, Genre, and Style.* Style DeKalb IL, 356.

Biswas, S. 2016. "Linguistic homophily in director networks and firm performance Linguistic homophily in director networks and firm performance." PhD thesis.

Brindle, A. 2009. "Just keep your pants on and in the closet and you'll be fine." The corpus analysis

of a white supremacist web forum". Presentation given at the Corpus Research Seminar, Lancaster University.

Boutyline, A., & Willer, R. 2015. "The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks."

Bowman-grieve, L. 2009. "Studies in Conflict & Terrorism Exploring "Stormfront": A Virtual Community of the Radical Right", *0731*(April). http://doi.org/10.1080/10576100903259951

Buckly, C & Salton, G. 1988. "Term-weighting approaches in automatic text retrieval". *Information Processing and Management*, 24 (5).

Buckley, C & Salton, G. 2016. "Onix Text Retrieval Toolkit." http://www.lextek.com/manuals/onix/stopwords2.html

Bunt, G. 2009. *iMuslims: Rewiring the House of Islam*. London: Hurst.

Burris, V., Smith, E. & Strahm, A. 2003. "White Supremacist Networks on the Internet." *Sociological Focus*, *33*(2).

Castillo, C., Mendoza, M., & Poblete, B. 2011. "Information credibility on twitter." *Proceedings of the 20th International Conference on World Wide Web* - WWW '11, 675. doi: 10.1145/1963405.1963500

Chen, H., Wang, F-Y. & Zeng, D. 2004. "Intelligence and Security Informatics for Homeland Security: Information, Communication, and Transportation*". IEEE Transactions on Intelligent Transportation Systems,* 5 (4), 329-341.

Chen, H. 2007. "Exploring Extremism and Terrorism on the Web : The Dark Web Project", 1–20.

Chertoff, M. 2008. "The Ideology of Terrorism: Radicalism Revisited*". Brown Journal of World Affairs*, 15 (1), 11-20.

Conover, M.D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., Menczer, F. 2011. "Political Polarization on Twitter." *in 'Proc. 5th International AAAI Conference on Weblogs and Social Media* (ICWSM)

Conway, M. 2002. "Reality Bytes: Cyberterrorism and Terrorist Use of the Internet."

Conway, M. 2016. "Studies in Conflict & Terrorism Determining the Role of the Internet in Violent Extremism and Terrorism: Six Suggestions for Progressing Research." *0731*(April). http://doi.org/10.1080/1057610X.2016.1157408

Coolsaet, R. 2011. *Jihadi Terrorism and the Radicalisation Challenge.* European and American Experiences, 2nd edition, pp. 269-287.

Cormen, T.H., Leiserson, C., Rivest, R & Stein, C. 2009. *Introduction to Algorithms* (3rd ed.). MIT Press and McGraw-Hill

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139164771

Danescu-Niculescu-Mizil, C. Gamon, M., Dumais, S. 2011. “Mark my words! Linguistic Style Accommodation in Social Media.” *WWW*, *78*(11), 1149. http://doi.org/10.1145/1963405.1963509

Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., & Kleinberg, J. 2012. “Echoes of power: Language effects and power differences in social interaction”. *Proceedings of the 21st International Conference on World Wide Web - WWW ’12*, 699. http://doi.org/10.1145/2187836.2187931

Davis, J. & M. Goadrich. 2006. “The Relationship Between Precision-Recall and ROC Curves.” *International conference on machine learning*, Pittsburg.

Duffy, M. E. 2003. “Web of Hate: A Fantasy Theme Analysis of Rhetorical Vision of Hate Groups Online”. *Journal of Communication Inquiry*, 27 (3), 291-312.

Easttom, C., & Taylor, J. 2011. *Computer crime, investigation, and the law. Course Technology*. Boston, MA, 15-28.

Gilbert, E. 2012. “Predicting tie strength in a new medium.” *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* - CSCW ’12, 1047. doi: 10.1145/2145204.2145360

Giles, H., Coupland, N. & Coupland, J. 1991. “Accommodation Theory: Communication, Context and Consequence”. *Contexts of Accommodation: Developments in Applied Sociolinguistics*, 1–68.

Gill, P., Emily, C., & Thornton, A. 2015. “What Are The Roles Of The Internet In Terrorism?”

Gillespie, T. 2012. *The Relevance of Algorithms.*

Graham, C. 2013. “Terrorism.com: Classifying Online Islamic Radicalism as a Cybercrime.”

Groh, G. & Hauffa, J. 2011. “Characterizing Social Relations Via NLP-based Sentiment Analysis.” *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 502–505.

Gruzd, A., & J. Roy. 2014. “Investigating Political Polarization on Twitter: A Canadian Perspective.” *Policy & Internet* 6 (1): 28–45.

Gustavson, A.T. & Sherkat, D.E. 2004. “Elucidating the Web of Hate: the Ideological Structuring of Network Ties Among Right Wing Hate Groups on the Internet.” *Annual Meetings of the American Sociological Association.*

van Gysel, C, M, de Rijke. 2015. “VOPE Datasets V1.” *VOX-Pol,* TNO.

Hara, K. O., & Stevens, D. 2015. “Echo Chambers and Online Radicalism: Assessing the Internet’s Complicity in Violent Extremism”, *7*(4), 401–422.

Han, B & Baldwin, T. 2011. “Lexical normalisation of short text messages: Makn sens a# twitter.” *In Proceedings of ACL*, volume 1.

Hankes, K. 2015. https://www.splcenter.org/hatewatch/2015/03/11/most-violently-racist-internet-content-isnt-stormfront-or-vnn-anymore

Hauffa, J., Lichtenberg, T., & Groh, G. 2014. “An evaluation of keyword extraction from online communication for the characterisation of social relations.” *Eprint arXiv:1402.2427*, 1–15.

Hoskins, A. 2011. *Radicalisation and Media: Connectivity and terrorism in the new media ecology.* Routledge.

Hutton, Christopher, M. 1999. *Fascism and the Science of Language* (Vol. 19). Routledge: London.

Jasnow, M., Crown, C. L., Feldstein, S., Taylor, L., Beebe, B., & Jaffe, J. 1988. “Coordinated Interpersonal Timing of Down-Syndrome and Nondelayed Infants with Their Mothers: Evidence for a Buffered Mechanism of Social Interaction.” *Biological Bulletin,* 175(3), 355. doi:10.2307/1541726

Koster, W. De, & Houtman, D. 2016. “Stormfront Is Like A Second Home To Me”, *4462*(April). http://doi.org/10.1080/13691180802266665

Kress, G. 1989. *Linguistics Processes in Sociocultural practice*. Oxford University Press.

Lanier, J. 2011. *You Are Not a Gadget: A Manifesto*. London: Penguin.

Lazer, D., Pentland, A., Adamic, L., Aral,S., Barabasi,A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara,T., King, G., Macy, M., Roy, D. & Van Alstyne, M. 2009. "Computational Social Science." *Science* 323: 721-723.

Margulies, Peter. 2004. “The Clear and Present Internet: Terrorism, Cyberspace, and the First Amendment.” *UCLA Journal of Law and Technology* 8(2).

Mcpherson, M., Smith-lovin, L., & Cook, J. M. 2001. “Birds Of A Feather : Homophily in Social Networks.”

Midlarsky, M. 2011. *Origins of political extremism. Mass violence in the twentieth century and beyond.* Cambridge: University Press.

Nichols, P.C. 1984. “Networks and Hierarchies: Language and Social Stratification”. *in Language and Power*. Sage Publications.

Paßmann, J., Boeschoten, T., & Schäfer, M. T. 2014. “The Gift of the Gab: Retweet Cartels and Gift Economies on Twitter.” *Twitter and Society*, 331–344.

Pedregosa. 2011. “Scikit-learn: Machine Learning in Python”, *JMLR 12*, pp. 2825-2830, 2011.

Pennebaker, J. W. 2011. “Your use of pronouns reveals your personality.” *Harvard Business Review*, *89*(December), 32–3. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22250353

Prentice, S., P., R., P., T. 2012. “The language of Islamic extremism: towards an automated identification of beliefs, motivations and justifications.” 1–39. http://doi.org/10.1075/ijcl.17.2.05

Rajalakshmi, R., & Aravindan, C. 2011. “Naive Bayes Approach for Website Classification.” 323–326.

Rheingold, H. 2000. *The Virtual Community*. MIT Press.

Rogers, R. 2004. *Information Politics on the Web.* MIT Press.

Rogers, R. 2013. *Digital Methods*. Cambridge: MIT Press.

Rusu, D. 2015. “Forum post classification to support forensic investigations of illegal trade on the Dark Web.” *TNO.*

Saever, N. 2014. *Knowing algorithms. Media in Transition*. Cambridge.

Schmid, A. P. 2013. “Counter-Radicalisation : A Conceptual Discussion and Literature Review.”

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Ungar, L. H. 2013. “Personality, gender, and age in the language of social media: the open-vocabulary approach.” *PloS One*, *8*(9), e73791. http://doi.org/10.1371/journal.pone.0073791

Sedgwick, M. 2010. “The Concept of Radicalisation as a Source of Confusion.” *Terrorism and Political Violence*, 22 (4).

Sproat, R., A.W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. 2001. “Normalization of non-standard words*.” Computer Speech & Language*, 15(3):287–333.

Sunstein, C.R. 2007. *Republic.com 2.0.* Princeton: Princeton University Press.

Statsoft.com. 2016. “Support Vectors (SVM)”. *http://www.statsoft. com/Textbook/Support-Vector-s*

Strathern, M. 2000. *Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy*. Routledge.

Thurlow, C. 2006. “From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media.” *Journal of Computer Mediated Communication,* 11, 667-701. http://dx.doi.org/10.1111/j.1083-6 101.2006.00031.

Whine, M. 1997. *Far Right on the Internet, Governance of Cyberspace.* Routledge: London

Yang, Y., & Eisenstein, J. 2011. “Putting Things in Context: Community-specific Embedding Projections for Sentiment Analysis.”

Zappavigna, M. 2011. “Ambient Affiliation: A Linguistic Perspective on Twitter.” *New Media Society,* 13, 788-806.

Zhao, D., & Rosson, M. B. 2009. “How and why people Twitter.” *In Proceedings of the ACM 2009 international conference on Supporting group work* - GROUP ’09 (p. 243). New York, New York, USA: ACM Press. doi:10.1145/1531674.1531710

Zhou, Y., Qin, J., Lai, G., Reid, E., & Chen, H. 2006. “Exploring the Dark Side of the Web: Collection and Analysis of U.S. Extremist Online Forums.” 621–626.

9. Appendix

The following sections include all scripts used for analysis and compiling of data during this thesis. It includes all scripts used for sampling, processing and analysing data. All processes are explaining within the lines of the scripts and signaled by a “#”.