An Integrated Framework For Detecting Online Harms, Modelling And Disrupting Of Cyberhate Networks

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

Wafa S. Alorainy

26th August 2022

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed . . . . (candidate)

Date . . . .

Statement 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed . . . . (candidate)

Date . . . .

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed . . . . (candidate)

Date . . . .

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed . . . . (candidate)

Date . . . .

To my parents, hope I made you proud
To my husband, for his love and support.

Abstract

Hate crimes are not a new phenomenon in society; however, social media and other means of online communication have begun to play an increasing role in hate crimes. Cyberhate, e.g. offensive or antagonistic language targeted at individuals and social groups based on their personal characteristics, which sometimes is considered a form of hate crime, is frequently posted and widely spread via the World Wide Web. The hateful individuals and groups who post this offensive or antagonistic language have increasingly been using the Internet to express their ideas and spread their beliefs. This spread facilitated by the Internet is considered a key risk factor for individual and societal tension leading to regional instability.

Automated Web-based cyberhate detection is important for observing and understanding community and regional societal tension - especially in online social networks where posts can be rapidly and widely viewed and disseminated. While previous work has involved using lexicons, bags-of-words, or probabilistic language parsing approaches, they often suffer from a similar issue, which is that cyberhate can be subtle and indirect (or implicit). Thus, depending on the occurrence of individual words or phrases, the analysis can lead to a significant number of false negatives, providing inaccurate representation of the trends in cyberhate. This problem was a motivation to challenge the thinking around the representation of subtle language use, such as references to perceived threats from "the other" including immigration or job prosperity in a hateful context. This thesis does this by proposing a novel "othering" feature set that utilises language use around the concept of "othering" and intergroup threat theory to identify these subtleties, implementing a wide range of classification methods using embedding learning to compute semantic distances between parts of speech considered to be part of an "othering" narrative. This novel feature resulted in a noticeable improvement for the classifier performance for both direct and indirect contextual hate.

In addition, understanding the network characteristics of these hateful groups could help to understand individuals' exposure to hate and derive intervention strategies to mitigate the dangers of such networks by disrupting communications. Concerning the people who post hateful content, this thesis analyses their hateful networks in order to build extensive knowledge of hateful group communication. This analysis shows that hateful networks exhibit higher connectivity characteristics when compared to other "risky" networks, which can be seen as a risk in terms of the likelihood of exposure to, and propagation of, cyberhate.

This thesis also examines several strategies for disturbing these risky networks. Results show that removing users with a high degree is most effective in reducing the hateful followers' network connectivity (GC, size and density) and, therefore, reducing the risk of exposure to cyberhate and stemming its propagation. This thesis further reveals that there are notable performance differences between these strategies and their effect on the disruption of hateful networks.

The experimental results demonstrated in this thesis contribute to the development of an integrated framework for the countering of cyberhate by proposing a novel feature set for detecting implicit cyberhate, analysing hateful networks and examining several network disrupting strategies.

Acknowledgements

First of all, praise and thanks be to God for granting me the capability and strength to successfully pursue my PhD. Thank you so much for giving me more than I could ask for and so, God, let the knowledge that you have granted me be useful for myself and others. Let it be an argument for me and not against me.

I am heartily grateful and thankful to my primary supervisor, Professor Pete Burnap for his uncountable encouragement, guidance and support from the beginning of my PhD. His professional excellence enabled me to develop my research skills and tackle the challenges of this research. Thanks for his great enthusiasm and patience with me. I would also like to thank my second supervisor Professor Matthew Williams for his knowledgeable advice and guidance. Also, I would like to thank Dr. Han Liu and Dr. Luca Giommoni for the wonderful discussions that enabled the research to move in the right direction. I would also like to thank the staff and PhD students of the School of Computer Science and Informatics, both past and present, for their support and feedback.

I wish to express my deepest gratitude to my dear family, Mother Mrs. Fatima and Father Mr. Suleiman, words cannot express how I am grateful to you, for your prays, support and love and may Allah reward you with health. Sisters and brothers, who always encourage and support me, thank you. I'm really blessed to have a wonderful and supportive family like you. I wish you all success in your careers. Also, thanks is due to my family-in-law; sisters, brothers and special thanks to mother-in-low Ms.

Sara, I remember your prays and best wishes and may Allah reward you with health.

My little angels, AlGhaniah and AlMuhra, who graciously accepted me leaving them for long hours as Mum has to work hard to finish important research, your presence inspired, touched and illuminated my heart during my PhD journey.

Finally, I owe thanks to a very special person, my friend and husband, Hamad, for his continued and unfailing love, support and understanding during my pursuit of PhD degree. Thank you for providing a welcome ear to listen to my thoughts during the hard times and tried to help by sharing a discussion. I greatly value his patience, contribution and deeply appreciate his belief in me. It is thanks to him I was able to come to the finish line of my PhD. Thank You!

Contents

List of Publications

The work introduced in this thesis is based on the following publications.

- Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L Williams. 'The Enemy Among Us': Detecting Cyberhate Speech With Threats-based Othering Language Embeddings. ACM Transactions on the Web (TWEB), 13(3):1-26, 2019 [27].
- Wafa Alorainy, Peter Burnap, Han Liu, Matthew Williams and Luca Giommoni. Disrupting networks of hate: Characterising hateful networks and removing critical nodes. Social Network Analysis and Mining, Article number: 27 (2022) [26].

List of Figures

List of Tables

List of Algorithms

List of Acronyms

DSRM	Data Science Research Methodology
API	Application Programming Interface
ITT	Intergroup Threat Theory
NLP	Natural Language Processing
BOW	Bag-Of-Words
CBOW	Continuous Bag-of-Words
TF-IDF	(Term Frequency-Inverse Document Frequenc
POS	Part Of Speech
TD	Typed Dependency
ML	Machine Learning
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
SVM	Support Vector Machines

NB Naive Bayes

AUC Area Under the Curve

DT Decision Tree

MLP Multilayer Perceptron

RF Random forest

LR Logistic Regression

PV-DBOW Paragraph Vector-Distributed Bag-Of-Words

PV-DM Paragraph Vector-Distributed Memory

NN Neural Network

RNN Recurrent Neural Networks

CNN Convolutional Neural Networks

GloVe Global Vectors for Word Representation

LSTM Long Short-Term Memory

DNN Deep Neural Networks

OSN Online Social Networks

SNA Social Network Analysis

GC Giant Component

ASP Average Shortest Path

PV-DM Paragraph Vector-Distributed Memory

Glossary

Cyberhate hate accrued using Internet.

Classification is the process of predicting the class of given data points.

Classifier is any algorithm that sorts data into labelled classes, or categories of information.

Algorithm a finite sequence of rigorous instructions, typically used to solve a class of specific problems or to perform a computation.

Precision is the fraction of relevant instances among the retrieved instances.

Recall is the fraction of relevant instances that were retrieved.

F-measure or F-score a measure that combines precision and recall is the harmonic mean of precision and recall

Dataset is a collection of data.

Social Network he use of internet-based social media platforms to stay connected with friends, family, or peers.

Graph a mathematical representation of a network and it describes the relationship between links (lines) and nodes (points).

Directed Network also called a directed graph, is a network in which the edges have a direction.

Metric is measures of quantitative assessment commonly used for comparing, and tracking performance or production.

Node a vertex or node is the fundamental unit of which graphs are formed.

Edge for a directed graph, the edge is an ordered pair of nodes. The terms 'arc,' 'branch,' 'line,' and 'link are sometimes used instead of edge.

Bridge a bridge, isthmus, cut-edge, or cut arc is an edge of a graph whose deletion increases the graph's number of connected components.

Clique a clique is a subset of vertices such that every two distinct vertices in the clique are adjacent.

Adjacent in a graph, two vertices are said to be adjacent, if there is an edge between the two vertices.

Scale Free Network is a network whose degree distribution follows a power law, at least asymptotically.

One-mode or Generic Network networks with one set of nodes that are similar to each other.

Two-mode or Bipartite Network a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V.

Removal Strategy strategy used for finding the most important node which if removed the network lost its connectivity.

Hybrid Removal Strategy removal of the most important nodes based on multiple removal strategies

Chapter 1

Introduction

1.1 Introduction

In recent years we have seen a transformation from ’tangible’ societies, where social interaction is mainly carried out face-to-face, into digital societies in the form of ’social media’ on the web [330]. The uptake of online social networks for social participation and social mobilisation is having a huge impact on society. While the benefit of online social media is enabling distributed societies to be connected, one disadvantage of the technology is the ability for hateful and antagonistic content, or cyberhate, to be published and propagated [331].

Hatred is not always expressed openly, but when it is it can take the form of subversive attempts to denigrate the target, verbal attacks or physical violence. According to Staub *et al.* [293], hate includes a negative evaluation of the object of hate: ’A hater sees the object of his or her hate as profoundly bad, immoral, dangerous, or all of these. The intense devaluation and the associated feelings make it satisfying to have the hated other suffer, experience loss, and be harmed’. Its expression can take the form of obviously, threatening and insulting comments, whether it be by tweet messages, visual images, Facebook messages or online YouTube videos, all of which can have a harmful effect on the victims who are targeted, and their families [24].

There are tragic examples of how hate speech online can spill over into the offline world, with horrific consequences, such as numerous occurrences of suicide attacks and the massacre of Muslims in New Zealand in 2019. In 2016 and 2017, the UK's decision to leave the European Union, and a string of terror attacks, were followed by noticeable and unprecedented increases in cyberhate [332], involving the rhetoric of invasion, threat and otherness [27].

Glaser *et al.* [133] suggested that racists often express their views more freely on the internet. Lee *et al.* [193] found that the implicit messages often used by haters communicating online were more persuasive to adolescents, who have become the target of new member recruitment of many hate groups. These adolescents might be easily influenced to conduct hate crimes.

The hateful content itself is often propagated by hateful groups. These hate groups have been increasingly using the internet to express their ideas, spread their beliefs, and recruit new members [193]. Zhou *et al.* [359] found that one of the major objectives of these websites was to share their hateful ideology. Some hate websites were associated with hate groups while others were maintained by individuals [128]. In addition to hateful websites, evidence shows that more instances of online hate speech occur on social media [144, 175, 287]. Evidence also suggests that the increase in exposure to online hate speech can be associated with the nature and impact of certain social events and conditions, e.g. terrorist events [61]. Since then, initial research interest in online hate has centred on the detection of hateful content and the characteristics of online hate groups [66, 107, 128]. This has attracted attention in exploring online hate on specific social media platforms [63, 224] and/or using computational methodologies to detect, remove, and understand the dynamics of hateful content distributed through these tools [229].

However, there are issues related to the methods intended to counter online hate, in particular relating to the detection of cyberhate and the reduction of the propagation of the hateful content. The following subsections illustrate these problems clearly.

1.1.1 Cyberhate Detection

Here, by **cyberhate** the author means explicit or implicit hateful content that is posted on the web targeting individuals or groups because of specific or identified characteristics such as their religion or ethnicity.

For the detection of cyberhate, it is important to recognise that expressing discriminative opinions, which are considered a form of online hate [309], involves different language uses. For example, words might be used to convey intense dislike such as 'hate them'; moreover, to encourage violence, an inflammatory verb could be used such as *"kill"*. While these examples contain directly threatening or offensive words (*kill, hate*), some examples include words that, on their own, would not constitute discriminative opinions (e.g., *send them home*). Although such messages do not contain explicitly hateful words, they are conveying the desire to distance different groups, within which there is inherent promotion of discrimination and division within society, fostering widespread societal tensions[1], [334, 167]. The context in which distancing terms are used is, of course, important here. For instance, if two boxers are fighting and an audience member shouts 'kill them', then this is a case of distancing and discrimination, but it is in a non-societal context.

There have been a number of attempts to automatically identify and quantify cyberhate by using different approaches, such as lexicons [131], syntactic [136] and semantic [183, 36] features - yet the limitation lies in classifying text that does not contain clear hateful words and would have an impact on classification accuracy, (e.g. *get them out of our country*). This was the motivation to propose a novel feature set that helps the machine to identify this sort of 'implicit' hate.

Recent studies have begun to interpret the effective features for machine classification of abusive language by focusing on how language is used to convey hateful or antagonistic sentiment. 'Othering' - the use of language to express divisive opinions between

[1] https://www.article19.org/wp-content/uploads/2018/06/UK-hate-speech_March-2018.pdf

the in-group ('us') and the out-group ('them') - has been identified as an effective feature for cyberhate detection[63].

On this point, this thesis aims to develop a novel method for cyberhate classification based around (i) the use of **two-sided pronouns** that combine the in-group and out-group (e.g. your/our, you/us, they/we), and (ii) the use of pronoun patterns, such as verb-pronoun combinations, which capture the context in which two-sided pronouns are used (e.g. send/them, protect/us). The author hypothesised that considering these linguistic features will provide an additional set of qualitative features that will improve classification performance. These features are subsequently used in combination with a paragraph embedding algorithm that infers semantic similarity between features to create a model that represents 'othering' language which is used for cyberhate classification.

1.1.2 Cyberhate Networks Characterisation

The data posted in online social media are not the only signals which may be used to study hate speech in Online Social Networks (OSN). Characterising hateful groups provides the benefits of detecting hateful content and presents plenty of opportunities to explore richer information related to hateful content exposure and propagation. Seeing other people post prejudiced comments online can lead to the adoption of an online group's biases and can influence an individual's own perceptions and feelings toward the targeted stigmatised group [152]. In addition, research on cyberhate also suggests that being exposed to hate speech can lead to an increase in outgroup prejudice toward groups targeted by such speech [290]. Thus, the problem of the presence of hateful groups relates to the exposure to and propagation of the hateful content itself. Studying the existing literature, it seems there is yet to be a study of *multiple* hateful networks with the aim of understanding whether there is evidence of similar 'levels of friendship', and therefore a general exposure to hate, or similar levels of propagation behaviour and therefore a general contagion effect. This research aims to address this gap in the knowledge by performing a baseline study that characterises several hateful networks extensively from multiple perspectives - namely, exposure to cyberhate (in followers' networks), and diffusion of cyberhate (in retweets' networks) by applying Social Network Analysis (SNA) methods to Twitter hateful networks. SNA is the process of investigating social structures through the use of networks and graph theory. It characterises networked structures in terms of nodes (user accounts) and the ties, edges, or links (relationships or interactions) that connect them.

1.1.3 Cyberhate Disruption

In addition to characterising hateful groups in terms of the exposure to and the propagation of their content, this work goes beyond this by experimenting with the disruption of content exposure and propagation as individuals and groups increasingly use the Internet to express their ideas, spread their beliefs, and recruit new members [193].

Also, online social media platforms are challenging to regulate [177], and policy-makers have struggled to suggest practicable ways of reducing cyberhate [7]. This is because removing certain content from a particular online source cannot guarantee the unavailability of the same content elsewhere [214]. In addition, it may be considered to go against freedom of expression. Thus, it is essential to study different strategies for disturbing the flow of cyberhate and reducing the exposure of others to it on Twitter.

This insight directly responds to the UK's Online Harms whitepaper, which focuses on the need to protect citizens online [339]. Disruption methods could include the possibility of identifying contagion pathways in hateful networks and evaluating the reduction in exposure of the network's users to receiving hateful content, in the same way as we might expect the spread of a traditional offline virus to be contained. Disruption methods include a wide variety of themes and approaches used to disrupt a network. For example, disruption methods may be applied to different network typologies or (modes) to better understand their effect. Most networks are defined as generic (or so-called one-mode) networks with one set of nodes that are similar to each other. However, several networks are, in fact, two-mode networks (also known as affiliation or bipartite networks; [56, 190]. These networks are a particular kind, with two different sets of nodes and ties existing only between nodes belonging to different sets. Typical examples include actor-by-event attendance and actor-by-group (or person) membership. In the real world, bipartite network communities have been found to solve many practical problems. For example, the study of bipartite network mining provides a new basis for research and a new method for the epidemic model [138]. By applying the disruption methods to bipartite networks, we can obtain deeper topology, hidden meaning and different network information, e.g bipartite analysis can detect groups/event links between nodes; therefore, it is possible that these groupings could identify collaborating/collusive subnetworks of actors creating and spreading hate.

Our study investigates the disruption strategies on the generic networks (one-mode network) where a tie connects two nodes (e.g. A follows B or retweet).

Also, we investigate the disruption strategies on the bipartite versions of hateful networks where two nodes in a set belong to the same relationship if they are both connected to the same node from the other set (A is connected to B by an affiliation relationship, e.g. a follow relationship).

Previous studies are yet to propose disrupting methods on multiple Twitter networks, specifically network node removal strategies, to prevent cyberhate from spreading. Therefore, this thesis addresses the lack of such a study by examining the disruption methods for the curtailing and containment of cyberhate (through network pruning). In particular, it aims to find a set of nodes whose removal from the network results in the fragmentation of the network into disjointed networks. Understanding how a network changes in response to node deletion is critical in many empirical networks as it reveals the most influential haters who have the greatest effect on hateful content propagation.

This thesis contributes towards the scientific understanding of detecting and managing cyberhate, advancing the integration of machine intelligence into this problem space. It does this by focusing on testing novel features for improving the process of detecting direct and indirect cyberhate. In addition, the study focuses on the analysis of individuals' exposure to online hate and individuals' role in propagating online hate amongst groups. Further, through targeted disruption in the flow of hate, this thesis concentrates on how cyberhate can be disrupted and how to reduce exposure to it.

1.2 Thesis aim and Objectives

This thesis aims to develop an integrated framework for detecting online harms, modelling and disrupting cyberhate networks. In the scope of this work, we have three interlinked objectives, which together help us understand, model and disrupt the flow of cyberhate:

- O1: To develop a classification model that takes advantage of the existence of the othering language in a hateful tweet to detect implicit and explicit cyberhate. This is discussed in Chapter 4.
- O2: To understand the characteristics of online hateful networks in a non-representative small sample, and model individuals' exposure to hate and cyberhate propagation. This is discussed in Chapter 5.
- O3: To deploy disruption methods, also called node removal strategies [160, 53], to curtail and contain cyberhate (through network pruning) on Twitter. This is discussed in Chapter 6.

1.3 Hypotheses and Research Questions

This thesis introduces three main hypotheses relating to these three objectives (O1-3):

- H1 (relating to O1): *linguistic features associated with othering language provide an additional set of qualitative features that improve the classification performance.*
- H2 (relating to O2): *hateful networks are similar in terms of online hate exposure and online hate propagation, and more connected than âriskâ networks.*
- H3 (relating to O3): *Applying node removal strategies (disruption strategies), depending on the node role in the network, reduce network connectivity (exposure reduction) and diffuse the spread of hate (contagion reduction).*

In order to develop evidence in support of these hypotheses, the following set of research questions was addressed:

- **H1:**

  RQ1: *To what extent can using othering and ITT theories drive the development of new features for classifying cyberhate and improve the performance of machine learning for cyberhate detection?*

- **H2:**

  RQ2: *By studying multiple hateful networks on Twitter, is there evidence of similar 'levels of friendship' across multiple hateful networks, and therefore a general measure of exposure to cyberhate?*

  RQ3: *By studying multiple hateful networks on Twitter, is there evidence of similar levels of propagation behaviour and, therefore, general contagion effect?*

- **H3**:

RQ4: *According to the structural characteristics of networks, which node removal strategies would be most effective at decreasing the propagation of hateful content?*

RQ5: *According to the structural characteristics of networks, would a combination of two (hybrid) node removal strategies be more effective at decreasing the propagation of hateful content - compared to applying a single node removal strategy?*

RQ6: *Would applying the node removal strategies on a bipartite version of the hateful networks improve the node removal strategies in terms of detecting the most important users?*

1.4 Contributions

- **C1:** The development of a novel ’othering’ feature set that utilises language use around the concept of ’othering’ and Intergroup Threat Theory to identify the subtleties of implicit cyberhate. A wide range of classification methods was implemented using embedding learning to compute semantic distances between parts of speech considered to be part of an ’othering’ narrative and improve the state-of-the-art embedding models in cyberhate detection by 2%-59%. When tested on unseen data using four different types of cyberhate relating to: religion; disability; race and sexual orientation, F-measures of 0.81, 0.71, 0.89 and 0.72 were obtained, respectively. Furthermore, the experiments show that different types of hate speech have different language characteristics and the use of othering terms can be effective for some but not all contexts of hate speech. This contribution addresses RQ1 and provides support for H1. The novel research has been published in ACM Transactions on the Web (TWEB) [27]

- **C2:** To the best of the author's knowledge,this is a brand-new investigative study carried out to understand the connectivity characteristics of two hateful follower networks. The analysis shows that the level of connectivity of the hateful followers' networks is similar, and therefore there are common levels of users' exposure to cyberhate. Hateful networks were also compared to another form of 'risky' network (i.e a suicidal ideation network of similar size) to understand the general level of the hateful networks' connectivity. The results showed evidence of higher connectivity between the hateful users (higher exposure to the hateful content) compared to the suicidal users, which suggests a potential virality of hateful content. Such users, however, have less reciprocated friendship behaviour than suicidal users (less connected around the topic). In addition to the contribution to knowledge, this study resulted in a fresh friendship datasets in the field that could be used in further research studies. This contribution addresses RQ2, and provides support for H2. It has been published in Social Network Analysis and Mining (SNAM).

- **C3:** To the best of the author's knowledge, this is this is a brand-new investigation carried out to understand the communication characteristics of three hateful retweets networks. Analysis shows several structural similarities were observed among the retweets' networks as were differences between the hateful retweet network. Also, there was a consistently and significantly greater reach of content (contagion), and greater degree of co-operation on the spread of the message (hate) in hateful networks, across all three hateful networks, and less in the comparator 'risky' network - suicidal ideation. Hateful content reaches more users in fewer hops. This contribution addresses RQ3 and provides additional support for H2.).

- **C4:** To the best of the author's knowledge, this is a novel study to develop strategies that identify nodes within hateful networks (user accounts) whose removal is empirically shown to reduce connectivity (largest component, density and average shortest path) in both the follower and retweet networks. Thirteen node-removal strategies, including a random-based strategy, based on network connectivity were tested on three network metrics: giant component size, density and the average shortest path. These strategies were applied to generic networks and bipartite networks. The experiments carried out for this study show that the best node-removal strategy is the degree-based strategy (single node removal) which has the highest impact on reducing the size of the largest component of the generic hateful followers' and retweets' networks. The rigour of these findings is demonstrated on two hateful followers' networks and three hateful retweets' networks. This contribution addresses RQs 4-6, and supports H3. This novel research also been published in Social Network Analysis and Mining (SNAM).

1.5 Thesis Structure

The outline for the remainder of this thesis is as follows:

- Chapter 2 - Background and Related Work - provides an overview of the processes related to detecting and managing online hate speech in social media. The chapter begins by analysing existing definitions of online hate speech and goes on to review the related literature on online hate detection by defining the fundamental concepts of the relevant methods for linguistic features, vector space embedding features, and machine learning. In addition, this chapter explores the fundamental concepts of Social Network Analysis (SNA) methods and network disruption strategies. The significant gaps for each concept are also summarised in this chapter.
- Chapter 3 - The primary purpose of this chapter is to clarifies the research design and research methodology. A general overview of the approach adopted in this thesis is given here. In addition, a description of the datasets collected is also given.

- Chapter 4 - Hateful Content Detection (RQ1) - introduces a novel method for cyberhate classification based on (i) the use of two-sided pronouns that combine the in-group and out-group (e.g., your/our, you/us, they/we) and (ii) the use of pronoun patterns such as verb-pronoun combinations, which capture the context in which two-sided pronouns are used (e.g., send/them, protect/us). A wide range of classification methods were implemented to compare our novel approach that fuses embedding learning with an 'othering' narrative to state-of-the-art methods. The chapter also qualitatively evaluates how the novel feature set can predict the vector space similarity between different kinds of hateful content. The associated chapter (4) makes up the contribution to C1. This contribution has been published [25].

- Chapter 5 - Networks of Hate (RQs 2 and 3) - provides an overview of a study related to hateful networks' structures. Several hateful networks are characterised extensively from multiple perspectives, namely: exposure to cyberhate (in follower networks) and diffusion of cyberhate (in retweet networks). A range of hateful networks is used to compare and contrast baseline measures of connectivity and propagation across multiple hateful networks. This associated chapter makes up the contribution to C2 and C3. This contribution has been published [26].

- Chapter 6 - Disrupting The Hateful Networks (RQs 4-6) - introduces node removal strategies and the effectiveness of removing nodes on reducing network connectivity (exposure reduction) and potentially diffusing hate (contagion reduction). This associated chapter makes up the contribution to C4. The answer of RQ4 has been published in [26].

- Chapter 7 - Conclusion and Future Work - concludes the thesis by summarising the contributions and findings of this research. Also, it introduces thesis implications and limitations as well as highlighting proposals for future work.

1.6 Summary

This chapter introduces the problem definition and the motivations that inspired this thesis. It also provides an overview of the thesis contributions and structure. Before discussing the main technical contributions of this thesis, the next chapter provides more detailed background information and positions the thesis in the context of existing work.

Chapter 2

Background and Literature Review

2.1 Introduction

To date, research interest in online hate has centred on the detection of hateful content and the characteristics of online hate groups [66, 107, 128]. Researchers have tended to concentrate on exploring online hate content on specific social media platforms, e.g. Twitter [63, 224] and/or using computational methodologies to detect, remove, and understand the dynamics of hateful content distribution [229]. However, there are issues remaining that relate to the methods intended to manage online hate speech, and these are introduced and discussed in this chapter. This chapter provides a comprehensive overview of research conducted in the field. In each section, general methods are first described and then works that adapt these methods to the classification of cyber hate, and the characterisation and disruption of hateful networks. Figure 2.1 gives an overview of the topics covered in this chapter.

The chapter begins by reviewing the literature methods and then analysing existing definitions of online hate speech, and then goes on to review related literature on online hate detection, propagation and prevention. The purpose of this chapter is twofold: firstly, to refine our understanding of hate speech; secondly, to provide an insight into the wider research area, which provides the basis for shaping the contributions of this thesis (focused on hate speech detection, propagation and prevention methods).

The chapter is divided into five sections. Section 2.2 clarifies the literature review

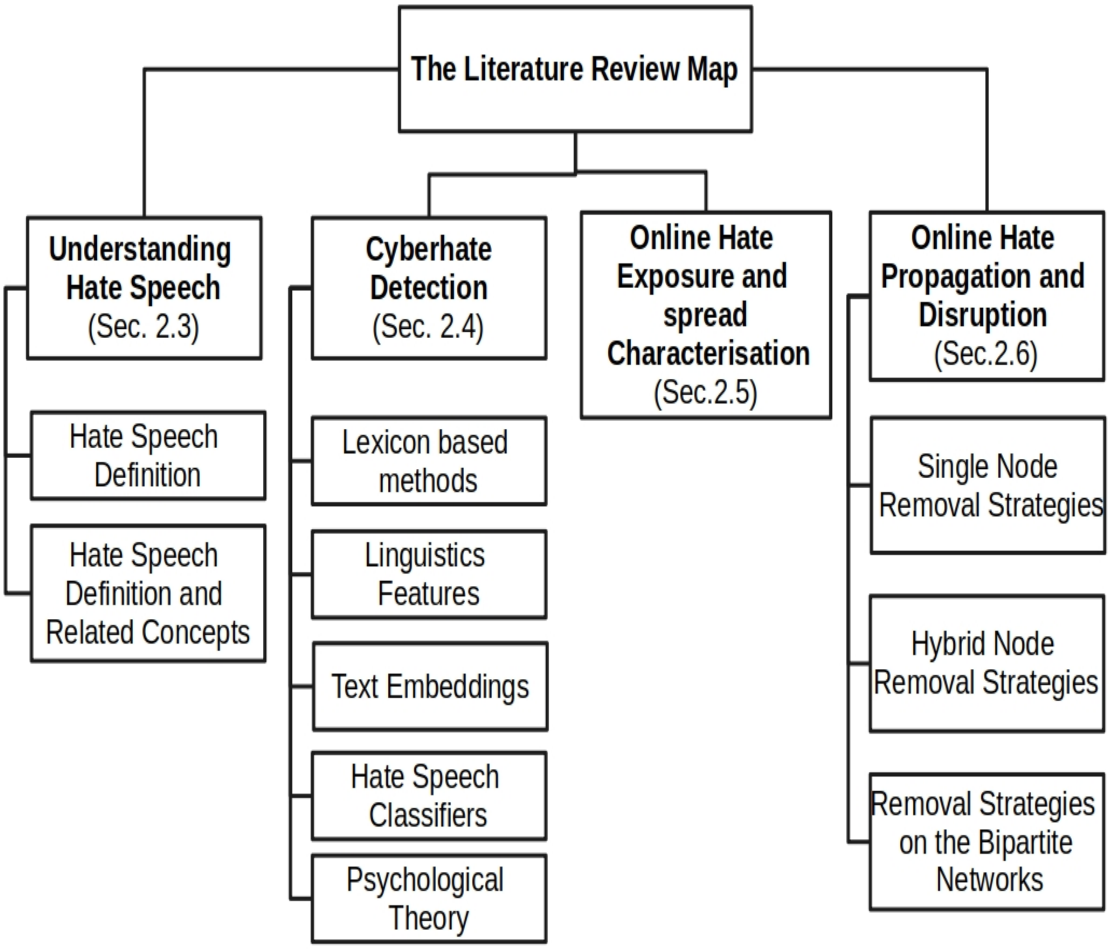

The Literature Review Map
Understanding Hate Speech (Sec. 2.3)
Hate Speech Definition
Hate Speech Definition and Related Concepts
Cyberhate Detection (Sec. 2.4)
Lexicon based methods
Linguistics Features
Text Embeddings
Hate Speech Classifiers
Psychological Theory
Online Hate Exposure and spread Characterisation (Sec.2.5)
Online Hate Propagation and Disruption (Sec.2.6)
Single Node Removal Strategies
Hybrid Node Removal Strategies
Removal Strategies on the Bipartite Networks

Figure 2.1: Road Map of Literature Review

methods. Section 2.3 explores existing approaches to defining online hate speech and provides a comparison of hate speech with other related concepts. This is followed by section 2.4, which presents an overview of the evolution of online hate detection in recent years, including methods of feature extraction and machine classification. Section 2.5 examines studies relating to the characterisation of the spread of hateful content on online platforms. Finally, hateful network prevention strategies are explored in Section 2.6. Within each section, the relevant literature is presented in tables set out in an analytical mode for further understanding of the methods related to this study. At the initial stage of this study, six open research questions are proposed and the answers to each research question are included, contributing to this thesis.

2.2 Literature Review Methods

For this process we applied which has been adapted by Fortuna *et el.* [118]. They presented a comprehensive survey with a critical overview on how the automatic detection of hate speech in text has evolved over the past years. Their method was chosen for this thesis because they claimed to have conducted a systematic literature review with the goal of collecting as many documents as possible in the field. The method is structured in four phases, that are presented and summarised in Figure 2.2. We describe here the different phases in more detail:

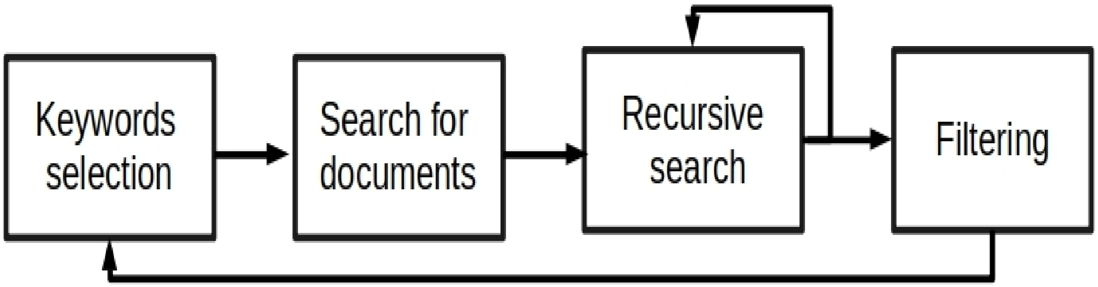


Figure 2.2: Literature Review Collection Methods

- **Keyword selection:** The first phase conducted was the keywords selection. We bear in mind that hate speech is a concept that has become more popular recently. Therefore, some other related concepts could have been used in the past by the scientific community [119]. The literature search relates to two areas: (i) searching for the general concept of the method used in the field and (ii) reviewing studies related to detecting, characterising and disrupting online cyberhate. To understand hate speech, we considered terms similar to hate speech like ’cyberhate’ and ’offensive’, ’implicit cyberhate’ and we also used terms referring to particular types of hate speech like ’sexism’ and ’racism’. For hate speech detection, we search for general concepts of features (e.g. lexicon meaning) used and classifiers performance (e.g. F-score meaning). In addition, we considered terms that refer specifically to the automatic detection of hate like ’cyberhate detection’, ’hate classification’ and ’hateful classifiers’. For online exposure and spread characterisation, we searched for general concepts of network characterisation (e.g. metrics), and we also used combined keywords like ’hateful networks’, ’online hate exposure’, ’hate spread’ and ’hate diffusion’ to review the related studies. For online hate disruption, we used keywords that refer to how to disrupt the network in general and then hateful networks specifically. Example keywords for the general concepts are: ’network attack’, ’networks disruption’ and ’networks dismantling’. Example keywords for reviewing studies related to hateful networks disruption are: ’hate disruption’ and ’terrorist network attacks’.

- **Search for documents:** We searched in different databases and services (ACM Digital Library, Scopus, Google Scholar, and DBLP), aiming to gather the largest possible number of documents in the areas of computer science and engineering.
- **Recursive search:** This step was to boost the search results. For each document which we felt was a useful source for our literature, we searched for the documents mentioned in the literature listed for that document. Also, we used Google Scholar to get both the references and documents that cited the original work. Recursively, we repeated the search with the new documents found. The search stopped at the point at which our work was initially accepted by the journal. As an example, our article which is related to hate speech detection was accepted by a TWEB journal at the beginning of 2018; therefore, the literature stopped on this date. However, we added all the recent studies into the appendix part of this thesis.

- **Filtering:** An initial step of filtering was conducted. Documents that were not related to computer science and not associated with different hate speech categories (general hate, cyberbullying, abusive, offensive, sexism, racism, etc.) studies were excluded. For example, some studies related to social and law were excluded.

2.3 Understanding Hate Speech

whilst the benefit of online social media platforms is their ability to facilitate connection between members of their own society and with members of other societies, a disadvantage of the technology is its ability to facilitate the widespread publication and propagation of hateful and antagonistic content *(hate speech)* [331]. This section explores definitions of hate speech and the nuances of the field's terminology.

2.3.1 Hate Speech Definition

Hate speech is defined in many ways, so it is important to understand the variety of definitions to improve the online hate detection process, for example, judging what is considered as hate (data annotation), forming features for automated hate detection, etc. There have been a considerable number of attempts to define hate speech as shown in the following Table 2.1 which illustrates the definitions with references.

Table 2.1: Hate speech definitions

Reference	**Hate definition**
Bansal *et al.* [38]	speech which contains an expression of hatred on the part of the speaker/author against a person or people based on their group identity.

Table 2.1 Continued: Hate speech definitions

ILGA [127]	Hate speech are public expressions which spread, incite, promote or justify hatred, discrimination or hostility toward a specific group. They contribute to a general climate of intolerance, which in turn makes attacks more probable against those given groups.
Nockleby ***et el.*** [244]	Hate speech is "usually thought to include communications of animosity or disparagement of an individual or a group on account of a group characteristic, such as race, color, national origin, sex, disability, religion, or sexual orientation.
Code of Conduct between EU and companies [10]	All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic.
Tarasova ***et al.*** [306]	Expression of hostility without any stated explanation for it
Nobata ***et al.*** [243]	Language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity
Twitter [4].	Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease

Table 2.1 Continued: Hate speech definitions

YouTube [3]	Hate speech refers to content that promotes violence or hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status and sexual orientation/gender identity. There is a fine line between what is and what is not considered to be hate speech. For instance, it is generally acceptable to criticise a nation-state, but not acceptable to post malicious hateful comments about a group of people solely based on their ethnicity
Facebook [5]	includes in their definition content that directly attacks people based on their: race; ethnicity; national origin; religious affiliation; sexual orientation; sex; gender or gender identity or serious disabilities or diseases
Mondal ***et al.*** [229].	Content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender or gender identity, sexual orientation, disability or disease is not allowed. However, clear attempts at humour or satire that might otherwise be considered a possible threat or attack are allowed. This includes content that many people may find to be in bad taste
Fortuna ***et al.*** [118]	Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.

Table 2.1 Continued: Hate speech definitions

The Encyclopaedia of the American Constitution [244]	Hate speech defines it as speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity
Gao *et al.* [126]	Explicit hate speech is easily identifiable by recognising a clearly hateful word or phrase. In contrast, implicit hate speech employs circumlocution, metaphor, or stereotypes to convey hatred of a particular group, in which hatefulness can be captured by understanding its overall compositional meanings
Benikova *et al.* [43]	expressing a (very) negative opinion against a target, they define explicit HS as expressing hateful sentiment, and implicit HS as the instances which do not express hateful sentiment, but a hateful stance.
De *et al.* [99]	Hate speech is a deliberate attack directed towards a specific group of people motivated by aspects of the group's identity.
Johnson *et al.* [166]	describe it as a type of speech that takes place online, generally social media or the internet, with the purpose of attacking a person or a group on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender.

Table 2.1 shows that there are similarities and differences between hate speech definitions. For example, all the quoted definitions, except [306], indicate that hate speech is directed at specific targets and it is based on particular characteristics of groups, such as ethnic origin, religion, or others. Several definitions use slightly different terms to describe when hate speech occurs. The majority of the definitions point out that hate

speech is designed to incite violence or hate toward a minority (e.g. ILGA, YouTube, Twitter). However, expressing hateful or discriminative opinions employs different language uses. For example, words might be used to convey intense dislike such as *'hate them'*; moreover, they may encourage violence; an inflammatory verb could be used such as 'kill'. While these examples contain directly/explicitly threatening or offensive words *(kill, hate)*, some examples contain words that, on their own, would not constitute hateful or discriminative opinions *(e.g. send them home)*, which is considered as implicit hate [126]. Notably, the explicitness/implicitness of the text is rarely mentioned in hate speech definitions. Yet, the implicitness of hate within a text is an important aspect to consider because it contributes to the mission of understanding and countering hate speech. There is a more nuanced side to hate speech which includes more implicit language, such as anti-religious, racist and anti-immigration sentiment [110]. This has an impact on communities and is important to take into account when detecting online hate. For this reason, the explicitness/implicitness of the text is an essential aspect and this research seeks to detect both the explicitness/implicitness of the text's (direct and indirect) hateful abuse. Table 2.2 shows that only two definitions mentioned the probability of hate content being directed implicitly or explicitly.

Explicit hatespeech is directed towards a group or individual based on protected characteristics, using a derogatory word or words that are clearly insulting, e.g. *'bastard'* or inciting e.g. *'Kill all Muslims'* [285, 126] or encouraging violence [1]. In contrast, **implicit hate speech** employs circumlocution, metaphor, or stereotypes to convey hatred of a particular group, in which hatefulness can be captured by understanding its overall compositional meanings [126]. Implicit hate speech is content that does not contain a clearly hateful expression, but one which implies hateful emotion toward a group or individual based on protected characteristics. An example of this content is 'bring them out', which, despite not containing any directly hateful words, implies a hateful emotion towards an expatriate person or group.

[1] https://dictionary.cambridge.org/us/dictionary/english/hate-speech

Table 2.2: Analysing the hate speech definition according to the focus of the hateful expression.

Definition reference	**Explicit**	**Implicit (indirect/subtle)**	**Humour**
Bansal ***et al.***[38]	✓	X	X
ILGA[127]	✓	X	X
Nockleby et el. [244]	✓	X	X
Code of Conduct between EU and companies[10]	✓	X	X
Tarasova ***et al.***[306]	✓	X	X
Nobata ***et al.***[243]	✓	X	X
Twitter[4]	✓	X	X
YouTube[3]	✓	X	X
Facebook [5]	✓	X	X
Mondal ***et al.***[229]	✓	X	X
Fortuna ***et al.***[118]	✓	X	✓
Encyclopedia of the American Constitution[244]	✓	X	X
Gao ***et al.***[126]	✓	✓	X
Benikova ***et al.***[43]	✓	✓	X
De ***et al.***[99]	✓	X	X
Johnson ***et al.*** [166]	✓	X	X

Several related studies frame these sorts of hate speech in their studies; for example, Waseem *et al.* [322] proposed a typology that synthesises the sub-tasks involved in hate speech detection. This includes directed hateful abuse or general abuse and further refines this into explicit or implicit hate. Also, Fortuna *et al.*[118] considers that all subtle forms of discrimination, even jokes, should be marked as hate speech. This is argued because this type of joke indicates relations between two groups: the jokers and those targeted by the jokes, relies on stereotyping and affects race relations [185]. Psychologists have shown that watching comedy which utilises prejudice as a device to get laughs can change behaviour for the worse. In one study, researchers found male test subjects were more reluctant to give a charitable donation and more willing to cut funding to a women's organisation after reading and hearing sexist jokes, but not neutral jokes [117]. In other studies, listening to sexist, anti-Muslim and anti-gay jokes increased acceptance of discrimination against women, Muslims and people of different sexual orientation [116]. The conclusion the psychologists came to was not that the comedy made the test subjects prejudiced, rather that the humour temporarily released them from having to regulate the prejudiced attitudes they already held, which they routinely suppressed due to wider societal pressure to appear non-prejudiced. Sexist, racist and homophobic jokes are a 'releaser' of prejudice that shapes behaviour [330]. The repetition of such jokes can act to reinforce racist attitudes [184] and, although they are often considered harmless, they can have a negative psychological effect on minority communities. The implicit aspect of hateful text is important because it is damaging to individuals and communities with protected characteristics, but the law doesn't include this type of hate as illegal content (unless it grossly offensive or inciteful of hatred), so it does not have to be removed. Yet, this implicit aspect needs to be monitored and observed to increase our understanding/management of the impact it has on certain communities. In the context of this thesis, we are interested in the explicit hate and the implicit hate towards people with protected characteristics. Indeed, all the definitions in Table 2.1 include pieces of valuable information and help us to understand the meaning of hate speech.Thus, hate speech is hereby understood to be inflammatory language that explicitly or implicitly targets an individual or group depending on protected characteristics .

2.3.2 Hate Speech Definition and Related Concepts

A further important aspect of the terminology is the confusion between the term hate speech and other terms such as 'cyberbullying', 'prejudice' and 'terrorism'. A way to better understand this complex phenomenon is by drawing a comparison with other related concepts. Focusing on these and understanding the differences and the similarities in the literature and empirical studies can provide an insight into how to select the studies which relate to hateful content/network detection. Table 2.3 illustrates the differences and the similarities between the definition used in this thesis and these terms.

Table 2.3: Comparison between hate speech definition and related concepts

Term	definition	Difference and similarity with the thesis perspective
Cyberhate	Cyberhate is a sort of hate speech, this term is for those who prefer to hold their conversations in 'cyber space'.	Cyberhate is considered as hate speech that exists in online media.
Terrorism	Any an abhorrent act of violence perceived as being directed against society- whether it involves the activities of anti-government dissidents, organised crime syndicates, common criminals, rioting mobs, people engaged in militant protest, individual psychotics, or lone extortionists- is often labelled as terrorism [149]	Hate speech could be considered as a type of terrorism [82] and can also follow an incident or trigger an event of terrorism.
Prejudice	An antipathy based upon a faulty and inflexible generalisation	In this study, negative prejudice toward people depending on their identity is considered a form of hate speech [174].

Table 2.3 Continued: Comparison between hate speech definition and related concepts

Extremism	Ideology associated with extremists or hate groups, promoting violence, often aiming to segment populations and reclaiming status, where outgroups are presented both as perpetrators or inferior populations [215]	This study considers the extremist as a producer of hate speech.
Radicalisation	Online radicalisation is similar to the extremism concept and has been studied within multiple topics and domains, such as terrorism, anti-black communities, or nationalism [11]	This study considers the possibility that radical people may use hate speech to spread their ideologies.
Abusive language	The term abusive language was used to refer to hurtful language and includes hate speech, derogatory language and also profanity [243]	This study considers hate speech as a type of abusive language

Table 2.3 Continued: Comparison between hate speech definition and related concepts

Cyberbullying	Aggressive and intentional acts carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who cannot easily defend him or herself [79]	Cyberbullying might be considered as hate speech if it is expressed in a hateful context and directed toward an individual or group based on the protected characteristics.
Discrimination	Process through which a difference is identified and then used as the basis of unfair treatment [309]	Hate speech is a form of discrimination through verbal means. Thus, this study's definition considers discrimination as a form of hate speech.
Flaming	Flaming are hostile, profane and intimidating comments that can disrupt participation in a community [137]	Flaming is considered in this study as hate speech when someone is being flamed based on protected characteristics.
Toxic language	Toxic language are rude, disrespectful or unreasonable messages that are likely to make a person to leave a discussion [8]	Not all toxic language contains hate speech or is directed toward individuals or groups based on protected characteristics; however, some hate speech does contain toxic language.

Table 2.3 Continued: Comparison between hate speech definition and related concepts

Offensive Language	Offensive language can include vulgar, pornographic, and hateful language. Vulgar language refers to coarse and rude expressions, which includes explicit and offensive reference to sex or bodily functions. Pornographic language refers to the portrayal of explicit sexual subject matter for the purposes of sexual arousal and erotic satisfaction. Hateful language includes any communication outside the law that disparages a person or a group on the basis of certain characteristics, such as: race; colour; ethnicity; gender; sexual orientation; nationality and religion [163].	Offensive language is included under the hate speech umbrella; this study considers the offensive language targeting individuals or groups based on protected characteristics as hate speech.

Table 2.3 Continued: Comparison between hate speech definition and related concepts

Harassment	conduct directed toward a victim that includes, but is not limited to, repeated or continuing unconsented contact that would cause a reasonable individual to suffer emotional distress, causing the victim to suffer emotional distress [111].	Harassment is a broad term and could be used in any context; however, it could manifest in online content. This study considers religious, race, gender and sexual orientation harassment as a sort of hate speech.
Aggressive	Overt, angry and often violent social interaction delivered via electronic means, with the intention of inflicting damage or other unpleasantness upon another individual or group of people, who perceive such acts as derogatory, harmful, or unwanted[151, 76].	Aggressiveness text is considered hate speech if related to protected characteristics.

In this work, both the explicit and implicit forms of online hate speech are referred to under the broad term **cyberhate** - which includes more subtle forms of hate. Understanding the nuance of hate speech definitions and terminologies helps refine methods for countering hate speech. The following section explores online hate detection methods, advantages and limitations.

2.4 Cyberhate Detection

Just as there is no clear consensus on the definition of hate speech, there is no consensus with regard to the most effective way to detect it across diverse platforms. As people increasingly communicate through Web-enabled applications, the need for high-accuracy automated cyberhate detection methods has also increased. Several studies have shown how individuals with biased or negative views towards a range of minority groups are taking to the Web to spread hateful messages [195, 263]. Instances of cyberhate and racist tension on social media have also been shown to be triggered by antecedent events, such as terrorist acts [65, 331]. This section reviews research carried out on automated cyberhate detection. In particular, related work aimed at detecting online hate speech, as well as work focused on sentiment or opinions that are deemed abusive [96]. The previous section of this work explained that online hate might be written explicitly or, implicitly, making the process of online hate detection more challenging.

This section is composed of five sub-sections: lexicon-based methods; linguistic features for cyberhate detection; the use of an embedding learning feature; cyberhate classification and exhibiting a psychological theory for the automated hate detection and othering language narrative surrounding cyberhate.

2.4.1 Lexicon-based Methods

Lexicon methods may involve the use of offensive words and slurs or negative/positive related words (e.g. emotions and negation words) as features, which might help to distinguish hate speech from other posts, yet they still have a weakness in terms of the ability to detect hate stereotypes when the text contains no single hateful words (implicit hate speech).

Lexicon-based methods also determine the sentiment or polarity of opinion via some function of opinion words in the document or the sentence [180, 153, 303]. The different techniques for a lexicon-based approach are a dictionary-based approach and corpus-based approach [271]. Dictionary-based approaches generally suffer from an inability to find offensive words with domain and context-specific orientations [96]. This method leads to the low recall problem [352] and poor precision [96] of the lexicon-based method, which depends entirely on the presence of negative/positive words to determine the sentiment's orientation.

Although one could say that these additional expressions can be added to the lexicon, such expressions change constantly, and new ones regularly appear following current trends and hateful events[352]. Moreover, this method does not have an effective mechanism for dealing with context-dependent opinion words. There are many such words; for example, the word 'black' can be a positive or negative word depending on the context[63]. There is probably no way to know the semantic orientation of a context-dependent opinion word by looking only at the word without prior knowledge of the entire context[105].

To overcome the disadvantages of the dictionary-based method, corpus-based approaches use a domain corpus to capture opinion words with preferred syntactic or co-occurrence patterns. Using the corpus-based approach alone to identify all opinion words is not as effective because it is hard to prepare a vast corpus to cover all words [140]. With this method, a lexicon is populated with words and phrases that are more attuned to the domain by incorporating contextual features that could possibly change the semantic orientation of an opinion word. However, features such as negations, like *'no, never, may'*, change the directionality of a lexicon item. For example, Gitari *et al.* [131] generated a lexicon of sentiment expressions using semantic and subjectivity features with an orientation towards hate speech, and then used these features to create a classifier for cyberhate detection. Similar to the previous study, Warner *et al.* [321] present a supervised approach that categorises hate speech by identifying stereotypes used in the text. However, their work depends on the existence of direct hate feature co-occurrence to decide the polarity of a specific tweet which might omit indirect/implicit hate speech.

Ding *et al.* [105] further explored the idea of intrasentential and inter-sentential sentiment consistency. Instead of finding domain-dependent opinion words, they showed that the same word might have different orientations in different contexts, even in the same domain. Silva *et al.* [287] detected cyberhate using sentence structure - specifically patterns starting with the word 'I'. They assume that the word 'I' means that the user is talking about emotions that he or she is feeling. Their work introduces the direction of the sentence structure rather than depending on specific words to recognise the sentence polarity. They suffered from a high false positive rate due to the model classifying sentences such as '*I hate following people*' as hateful.

Another direction for building lexicons is to follow a sentiment scoring method which uses emoticons, modifiers, negations and domain-specific words [32]. Despite the scoring method outperforming the baseline methods, it requires a manual scoring of words.

2.4.2 Linguistic Features

In general, while previous studies have addressed the difficulty in defining hateful language, their experiments have led to better results when combining a large set of linguistic features. A review of these results follows.

Similar to the use of dictionaries is the **bag-of-words** (BoW); it is one of the most basic forms of natural language processing based on the extraction approach [63, 136, 187]. BoW is a method that creates a vocabulary of all the unique words occurring in all the documents in a training set [252]. BoW has been successfully applied as a feature extraction method for the automated detection of hate speech, relying largely on keywords relating to offence and antagonism [252, 326, 62]. In this case, a corpus is created based on the words that are in the training data, instead of a predefined set of words, as in the dictionaries. After collecting all the words, the frequency of each one is used as a feature for training a classifier. The disadvantage of this kind of approach is that the word sequence is ignored, as is its syntactic and semantic content. Therefore, it can lead to misclassification if the words are used in different contexts. To overcome this limitation, n-grams can be adopted [130, 45].

N-grams are one of the most popular techniques in hate speech detection [36, 63, 96, 136, 243, 323]. The most common n-grams approach consists of combining sequential words into lists with size N. In this case, the aim is to enumerate all the expressions of size N and count all the occurrences. This improves classifier performance because it incorporates, to some degree, the context of each word. Instead of using words, it is also possible to use n-grams with characters or syllables. This method is not so susceptible to spelling variations, as when words are used. Character n-gram features have proved to be more predictive than token n-gram features for the specific problem of abusive language detection [217]. Use of a character n-gram-based approach outperforms word n-grams due to character n-gram matrices being far less sparse than the word n-gram matrices [323], [217]. Character n-grams have been shown to be more effective if joined with additional linguistic features including gender and location [323]. However, using n-grams also has drawbacks. One disadvantage is that related words can be a significant distance apart from each other in a sentence [63] and a solution for this problem, such as increasing the N value, slows down the processing speed [78]. Also, studies point out that higher N values (5) perform better than lower values (unigrams and trigrams) [204].

TF-IDF The TF-IDF (term frequency-inverse document frequency) has also been used in cyberhate classification problems [104, 36]. TF-IDF is a measure of the importance of a word in a document within a corpus and increases in proportion to the number of times a word appears in the document. However, it is distinct from a bag of words, or n-grams, because the frequency of the term is offset by the frequency of the word in the corpus, which compensates for the fact that some words appear more frequently in general (e.g. stop words) [269].

Part-of-speech (POS) indicates how the word functions in meaning as well as grammatically within the sentence. The part of speech approach makes it possible to improve the importance of context and detect the role of the word in the context of a sentence. It operates by detecting the category of the word, for instance, personal pronoun (PRP), Verb non-3rd person singular present form (VBP), Adjectives (JJ), Determiners (DT), Verb base forms (VB). Part-of-speech has also been used in hate speech detection [136]. With these features, it was possible to identify frequent bigram pairs, namely PRPVBP, JJDT and VBPRP, such as ’you are’ [104]. It was also used to detect sentences such as ’send them home’, ’get them out’ or ’should be hung’ [65]. However, POS caused confusion in the identification of classes when used alone as features [118].

Dirichlet Allocation (LDA) is a probabilistic topic modelling method. It is mainly used to give an estimation of the latent topics in a data set and these latent topics are then used as features instead of words. However, LDA is suitable for unsupervised and semi-supervised machine learning settings. Xiang et al. [342] claimed that BoW did not work well for abusive text detection on Twitter. Instead, they included highly expressive topical features and other lexicon features by using the LDA model. This approach can be an alternative for supervised methods. In addition [12] showed that topic modelling linguistic features were used to identify posts belonging to a defined topic (Race or Religion).

Typed dependencies is an approach that parses a sentence and represents its grammatical structure by defining the relationships between ’head’ words, and words that modify those heads[100]. Typed dependencies have been widely used for extracting the functional role of context words for sentiment classification [179, 153] and document polarity [304].

To understand the types of features that can be obtained with this approach, the Stanford typed dependencies representation provides a description of the grammatical relationships in a sentence that can be used by people without linguistic expertise [100]. This was used to extract Theme-based Grammatical Patterns [131] and also to detect hate speech and specific othering language [65, 62]. Some studies reported significant performance improvements in hate speech automatic detection based on this feature

[63, 131], for instance by reducing the false negative rate by 7%, beyond the use of BoW and known hateful terms [62, 63]. In many works, n-gram features are combined with other features. For example, Nobata *et al.*[243] reported that while token and character n-gram features are the most predictive single features in their experiments, combining them with all the additional features further improves performance. They also speculated that engineering features based on deeper linguistic representations (e.g. dependencies and parse tree) may improve classification results for contents on social media, which was shown by [63].

2.4.3 Text Embedding

A bag-of-words (BoW) representation can be seen as a very high-dimensional vector representation. An embedding is a translation of a high-dimensional vector into a low-dimensional space (e.g. 300 dimensions). Unlike in BoW representations, the unique dimensions in the vector space embedding typically have no specific meaning. Ideally, an embedding captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space. Text embedding learning is aimed at training a model that can automatically transform a sentence/word into a vector that encodes its semantic meaning. Embedding applications have been shown to be capable of capturing specific semantic features from complex natural language (e.g. location [256], entity [286] and images feature [18]). Several models, including neural net language models (NNLM), global vectors for word representation (GloVe), deep contextualised word representations (ELMo), FastText, Word2vec and Paragraph2vec, which are designed to learn word embeddings. It has been shown that embedding representation is very capable of semantic learning when word vectors are mapped into a vector space, such that distributed representations of sentences and documents with semantically similar words have similar vector representations [219] [220].

Based on the distributional representation of the text, several methods of deriving word representations that are related to cyberhate and offensive language detection are explored such as the FastText application[36], GloVe[262], Word2vec [219] and paragraph2vec [220].

FasText is a library for learning word embeddings and text classification created by Facebook's AI Research lab. The model allows an unsupervised learning or supervised learning algorithm to be created to obtain vector representations for words.

Also, **GloVe** is an unsupervised learning algorithm developed by Stanford to generate word embeddings by aggregating the global word-word co-occurrence matrix from a corpus.

Word2vec model involves word vector representations where conceptually, if two words are similar, they should have similar values in this projected vector space. Word2vec has two architectures: skip-gram and CBoW. While skip-gram aims to predict nearby words from a given word, CBoW predicts a target word from its set of context words.

In the **Paragraph2vec** model, paragraphs are represented as low-dimensional vectors and are jointly learned with distributed vector representations of tokens using a distributed memory model [220]. In the literature, paragraph2vec is also called sentences2vec, document2vec [220] and comment2vec [106]. Figure 2.3 show frameworks for learning word vector (left side) and paragraph vector (right side).

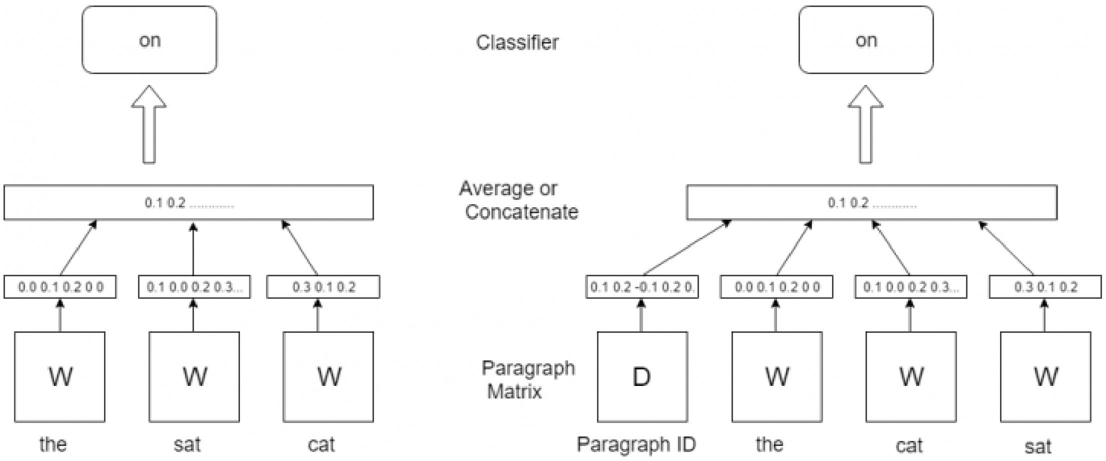


Figure 2.3: Word Embeddings Machine Learning Frameworks: word2vec and paragraph2vec.

Learning the paragraph embedding could be achieved using two models: (i) Paragraph Vector-Distributed Memory (PV-DM) or (ii) Paragraph Vector-Distributed Bag-of-words (PV-DBoW). The Paragraph Vector-Distributed Memory (PV-DM) model is similar to the Continuous-Bag-of-Words (CBoW) model in word2vec which attempts to predict the output (target word) from its neighbouring words (context words) with the addition of a paragraph ID. On the other hand, the DBoW model is similar to the skip-gram model of word2vec, which predicts the context words from a target word. Comparing the GloVe model to the Word2vec model, both models are word embedding models and they are predictive; the word2vec model only takes into consideration the local context; hence, it does not capture the global context. Although FastText predicts the unknown words, breaking words into several sub-words at the training time, it takes longer to train a FastText model compared to a Word2vec model.

In the existing literature, Schmidt *et al.* [281] identified that hate speech detection required sentence-level - rather than word-level - classification. Hence, some authors proposed sentence embeddings or comment/paragraph embeddings to solve this problem. For example, Djuric *et al.*[106] and Nobata *et al.*[243] presented the paragraph2vec approach to classify language in users' comments. These authors showed that using the vector space features, particularly, for a sentence (paragraph) embedding, improved the classifier performance compared to not applying vector space features. Specifically, Djuric *et al.* [106] implemented an extended version of Word2Vec for sentences which has been shown to outperform the BoW representation for cyberhate classification models by around 3% to 4% in F1 score. This suggests that paragraph vector features can be powerful when combined with standard NLP features [243].

2.4.4 Hate speech classifiers

Machine learning models take samples of labelled text to produce a classifier that is able to detect hate speech. They are trained using labelled data, annotated by content reviewers. Various models have been proposed and have proved successful in the past. A selection of open-sourced systems presented in the recent research follows. The methods utilised for hate speech detection in terms of classifiers are predominantly supervised learning approaches. As classifiers, the most widely used methods for hate speech detection were summarised in a survey by Fortuna *et al.* [118] and Macavaney *et al.* [208]. These include Naive Bayes (NB)[178] [11][65][36] Support Vector Machine (SVM), Logistic Regression (LR) [106][96], Random forest (RF), and Decision Tree (DT) [36][65][11]. Neural Network (NN) classifiers, e.g. Multilayer Perceptron (MLP) [119], DNN and CNN [36] were also used but are less popular. These models are commonly used in text categorisation. Naive Bayes models label probabilities directly with the assumption that the features do not interact with one another. Support Vector Machines (SVM) and Logistic Regression are linear classifiers that predict classes based on a combination of scores for each feature. The Random Forest (RF) classifier creates decision trees based on data samples, obtains the prediction from each of them, and finally selects the best solution by means of voting. Random Forest is a tree-based machine learning algorithm that leverages the power of multiple decision trees to make decisions[305].

Broadly speaking, Neural Networks (NNs) are inspired by how the human brain works. NNs are based on a collection of connected nodes called artificial neurons, which loosely model the neurons in the brain. Each connection, like the synapses in a brain, can transmit a signal between neurons. An artificial neuron that receives a signal can process it and send it to other artificial neurons to which it is connected. Examples of NN architectures used for classification of hate speech include Multilayer Perceptron (MLP). MLP is a feed-forward artificial neural network model that maps input datasets to an appropriate set of outputs. It is characterised by several layers of input nodes connected as a directed graph between the input and output layers [237]. MLP can provide competitive results on sentiment classification and factoid question answering [159].

Other NNs utilised include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) - specifically the Long Short-Term Memory network (LSTM) [83]. They treat the text as a sequence of words, and thus implicitly assume that at least some aspects of word order are important. In the context of hate speech classification, intuitively, CNN extracts word or character combinations [125, 36, 257], (e.g. phrases, n-grams). RNNs analyse a text word by word and store a representation of the already processed text as a fixed-dimensional vector in a hidden layer [188, 217]. RNN can learn word or character dependencies (order information) in tweets [36, 101]. Such methods, including deep neural networks (DNN), can be used to extract word or text embeddings as features, which are subsequently combined with another classifier (e.g. SVM) to use such embeddings as features for classification [106, 217].

What follows next is a review of the state-of-the-art features and classifiers used to detect cyberhate and a discussion of their strengths and limitations. To compare and contrast the state-of-the-art classifiers, it is necessary to first define how such methods are typically evaluated, i.e. the classifier's capability to predict the correct category, and not its computational complexity [284]. Ordinarily, the evaluation measures in classification problems are determined from a matrix with the numbers of examples correctly and incorrectly classified for each class, called a 'confusion matrix'. An example of a confusion matrix for a binary classification problem (which has only two

Table 2.4: Confusion matrix for binary classification.

	Predicted Class	
Actual class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

classes - positive and negative) is presented in Table 2.4.

TP (true positives) is the number of examples that are correctly predicted as belonging to the positive class. TN (true negatives) is the number of examples that are correctly predicted as belonging to the negative class. FP (false positives) is the number of examples that are classified as positive while they are from the negative class and FN (false negatives) is the number of examples that are classified as negative when their true class is positive. Common metrics for evaluating classification tasks using these figures include Accuracy, Precision, Recall, F1-score [227]and Area Under the Curve (AUC) [157], which are defined below:

$$Accuracy = \frac{(TP + TN)}{TP + FP + TN + FN} \tag{2.1}$$

$$Precision = \frac{(TP)}{TP + FP} \tag{2.2}$$

$$Recall = \frac{(TP)}{TP + FN} \tag{2.3}$$

$$F1 - score = 2 \cdot \frac{(Precision \cdot Recall)}{Precision + Recall} \tag{2.4}$$

$$AUC = \frac{1}{2} \cdot \left((\frac{TP}{TP + FN}) + (\frac{TN}{TN + FP}) \right) \tag{2.5}$$

Accuracy is the proportion of the correctly classified examples (i.e. true positive and true negative examples); precision measures the proportion of false positives; recall measures the proportion of false negatives, whilst the F1-score is the harmonic means of precision and recall. In addition, AUC provides an aggregate measure of performance across all possible classification thresholds, ranging from 0 to 1, a model with predictions that are 100% incorrect has an AUC of 0.0, and one with predictions that are 100% correct has an AUC of 1.0. Due to the often highly imbalanced number of positive vs. negative examples in binary classification, the negative class usually dominates the accuracy of a model, leading to misinterpretation of the results. For example, when the positive examples of a class represent only 1% of the test set, a trivial classifier that makes negative foresight for all examples has an accuracy of 99%. However, such a system is inefficient. For this purpose, precision, recall and F1-score are more commonly used instead of accuracy and (AUC) [141] for evaluating unbalanced classification problems.

Table 2.5 shows studies that relate to online hate classification. For each study, the table illustrates the classifier/s, the feature/s, the accuracy, the precision, F-score and the area under the curve (AUC).

Table 2.5: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Study	**Hate Type**	**Year**	**class ifier**	**Features**	**Accu-racy**	**Precis ion**	**Rec-all**	**F-score**	**AUC**
Greevy *et al.*[136]	Racism	2004	SVM	BoW, N-grams, POS	-	0.90	0.90	0.90	0.90

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Warner *et al.*[321]	Anti religious	2012	SVM	Template based strategies, word sense, disambiguation	-	0.68	0.6	0.63	-
Kwok *et al.*[187]	Aggression	2013	Naive Bayes	N-grams	0.76	-	-	-	-
Burnap *et al.*[65]	Anti religious	2014	Random Forest Decision Tree, SVM	N-gram, typed dependencies	-	0.89	0.69	0.77	-
Liu *et al.*[204]	Violence	2014	Naive Bayes	T-IDF, N-grams, topic similarity, sentiment analysis	-	0.97	0.82	-	-

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Djuric *et al.*[106]	Hate speech	2015	Logistic Regression	Paragraph embedding	-	-	-	-	0.8
Gitari *et al.* [131]	Ethnicity, religion and nationality	2015	Non-supervised	Rule-based approach, sentiment analysis, typed dependencies	-	0.65	0.64	0.65	-

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Burnap *et al.*[62]	Anti religious	2015	Random Forest Decision Tree, SVM, Bayesian Logistic Regression, Ensemble	N-gram, reduced typed dependencieshateful terms	-	0.89	0.69	0.77	
Nobata *et al.*[243]	Hate speech, profanity and derogatory language	2016	Skip-bigram Model	N-grams, length, punctuation, POS		0.83	0.83	0.83	

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Waseem *et al.*[323]	Sexism and racism	2016	Logistic Regression	User features	-	0.72	0.77	0.73	-
Burnap *et al.*[63]	Anti religious, racism, disability and sexual orientation	2016	SVM, Random Forest, Decision Tree	BoW, dictionary, typed dependencies	-	0.79	0.59	0.68	-

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Agarwal *et al.*[11]	Online radicalisation	2016	One-class Classifiers, Random Forest, Naive Bayes, Decision Trees	Topic modelling, sentiment analysis, tone analysis, semantic analysis, contextual metadata	-	0.73	0.86	-	-

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Fortuna *et al.*[119]	sexism, body, origin, other Lifestyle, racism, homophobia, religion and ideology	2017	MLP	n-grams	0.77	0.78	0.72	0.76	-
Badjatiya *et al.* [36]	**Sexism and racism**	**2017**	**GBDT**	**DNN (LSTM) Paragraph embedding**	**0.93**	**0.93**	**0.93**	-	-

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Davidson et al[96]	Hate speech and offensive	2017	logistic regression	bigram, unigram, trigram each weighted by its TF-IDF, binary and count indicators for hashtags, mentions, retweets, and URLs, the number of characters, words, and syllables in each tweet.	-	0.90	0.90	0.90	-

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Del ***et al.***[101]	Cyber bullism, incitement to self-harm practices and sexual orientation	2017	LSTM	POS, sentiment analysis, word2vec, CBoW, N-grams, text features	-	0.833	0.872	0.851	-
Gamback ***et al.***[125]	Sexism and racism	2017	Word2vec Model	character 4-grams, word vectors	-	-	-	0.78	-
Jha ***et al.***[165]	Sexism	2017	FastText	bag of words and bag of n-grams	-	-	-	0.87	-

Table 2.5 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Koffer *et al.*[183]	Refugee crisis	2018	LR	BoW, 2-grams, 3-grams, linguistics, word embedding, paragraph embedding extended 2-grams and extended 3-grams	0.70	-	-	0.70	-
Zhang *et al.*[355]	Religion and refugees	2018	CNN with gated recurrent unit (GRU) layers	Word embedding	-	-	-	0.71	-

BoW is used in [62, 136] with SVM and RF classifiers; however, this work suffers from a high rate of false positives, since the presence of hateful words can lead to the misclassification of tweets being hateful when they are used in a different context (e.g. 'black') [136]. In [243, 79, 323], n-grams have been used with SVM and LR classifiers to improve the performance of hate speech classification by capturing context within a sentence that is lost in the BoW model. However, the fact that not all content in a hateful community is hateful (i.e. it is used as a form of in-group affection) may mean this approach leads to false positives. The use of n-gram in combination with tf-idf features in [204] with the NB classifier, and combining the n-gram feature with typed dependencies in [65] using SVM and RF classifiers resulted in low recall. This is possibly because these types of combinations make no use of semantic similarities between words because they assume that the counts of different words provide independent evidence of similarity [147]. Using NN classifiers, MLP classifiers in combination with n-grams in [119] showed less promising results when compared to combining SVM classifiers with n-grams in [136]; however, Fortuna *et al.* [119] observed that the MLP classifier achieved better performance in detecting higher hateful samples compared to SVM, LR and RF when combined with n-gram features.

LSTM and FastText classifiers in combination with BoW and n-grams feature in [101] and Jha *et al.* [165] offered an improvement for hate classification compared to the non-neural network classifiers. In contrast, in a study by Zhang *et al.* [355], the CNN classifier was combined with word embedding features and resulted in no improvement compared to other classifiers. RNN and CNN can be computationally expensive compared to MLP and tend to be challenging to configure and require careful fine-tuning of hyper-parameters and the learned models are often difficult to interpret [222, 280]. Badjatiya *et al.* [36] compared the classification accuracy of a combination of different baselines and classifiers (Char n-gram, TF-IDF, BoW and LSTM) and found that learning embedding with gradient-boosted decision trees led to the best classification performance by 18% over state-of-the-art char/word n-gram methods. For German language processing, Koffer *et al.* [183] examined different types of features (BoW, 2-grams, 3-grams, linguistics, Word2Vec, Paragraph2vec, extended 2-grams and extended 3-grams) for training logistic regression LR classifiers. The experimental results, obtained based on a 75/25 split between the training and test data, showed that the best performing types of features are Word2Vec and Extended 2-grams.

The highest 'overall' f-score in Table 2.5 was obtained by the study by Davidson *et al.*[96], who classified text into racism, sexism or neither. Comparing their study with studies that used only word vector feature space, Gamback *et al.*[125] and Zhang *et al.* [355] show the advantage that was obtained by applying the paragraph vector space feature. However, they did not provide details of the detected hateful samples' f-scores. Knowing the f-scores of each class label provides an insight into the advantage of the feature sets. Although current methods have reported promising results, it is apparent that their evaluations are not generalised on an external dataset. Moreover, in these studies there is a little focus on classifying text that does not contain clearly hateful words (implicit hate) and this would have an impact on classification accuracy, *(e.g. get them out).*

While previous studies highlight the utility of methods capable of measuring semantic distances between words, such as embedding learning using individual words [106], and n-grams [243], the examples of implicit hate require an additional layer of qualitative context that sits above combinations of individual words. Recent studies have begun to interpret the effective features for machine classification of abusive language by focusing on how language is used to convey hateful or antagonistic sentiment.

Recent studies have suggested that utilising psychological theories for detecting cyberhate would contribute to improving the detection process because, ultimately, we are dealing with human expressions, even if they exist virtually [63]. For example, 'Othering' - the use of language to express divisive opinions between the in-group ('us') and the out-group ('them') - has been identified as an effective feature [63]. The concept of 'othering' offers a potential candidate framework for the aforementioned qualitative layer capable of capturing the more subtle expressions of cyberhate such as the 'send them home' example. Anti-Hispanic speech might make a reference to border crossing or crime, anti-African American speech often references unemployment or single parent upbringing, and anti-Semitic language often refers to money, banking and the media.

The use of stereotypes also means that certain types of language may be regarded as hateful Even if, when taken in isolation, no single word in the passage is hateful [123]. This raises the question of whether the incorporation of 'othering' into computational feature processing could further improve classification performance. The following section reviews the wider theories of othering and Intergroup Threat Theory (ITT), in the context of how it may be used for the detection of cyberhate, leading to the formal definition of a research question for the thesis.

2.4.5 Psychological theory: Othering language and Intergroup Threat Theory (ITT)

Psychological concepts that relate to defining prejudice, which is one form of hate speech, could expand our understanding of how to detect it online, particularly in the context of intergroup conflict and violence. It is beneficial to deliberate hate as a social as well as an individual phenomenon. There have been several attempts to use psychological theories to understand hate content online. Intergroup Threat Theory (ITT) and Otherness are examples of psychological models that incorporate similarity and intergroup conflict and, therefore, could be useful for greater understanding of cyberhate.

ITT posits that prejudice is a product of perceived realistic and symbolic threats. Realistic threats can be conceptualised in economic, physical and political terms[295]. Such threats refer to competition over material economic group interests, including scarce resources such as jobs, houses, benefits and healthcare, which may be embedded into online text containing subtle and implicit hate. Symbolic threats are based on perceived group differences in values, norms and beliefs. Out-groups that have a different viewpoint can be seen as threatening the cultural identity of the in-group [296]. Studies show that perceived threats to in-group values by immigrants and minorities are related to more negative attitudes towards these groups, unless countered by other in-group members [233]. For instance, research using ITT has recently focused on the perception of a threat from Muslims in Europe [89]. This can result in 'othering' language, such as 'get them out', which represents a speech act that aims to protect resources for the in-group. The core concept is that these resources and values are threatened by the out-group, leading to anxiety and uncertainty in the in-group [297]. The desire to protect the in-group is considered the underlying motivation responsible for negative attitudes and discriminatory behaviour.

Othering is an established construct in rhetorical narrative surrounding hate speech [216]. Likewise, there are contemporary uses of the concept of 'Otherness'; for instance, Andersson *et al.*[30] uses the concept in relation to radicalisation processes that affect first-generation Europeans. Lister *et al.*[201] defines othering as a 'process of differentiation and demarcation by which the line is drawn between "us" and "them" - between the more and the less powerful, and through which social distance is established and maintained'. Also, the 'we-they' dichotomy has previously been identified in racist discourse [338].

Othering has been used as a framework for analysing racist discourse from a qualitative perspective in previous work. For instance, Wodak *et al.*[337] argued that while the 'self' or the concept of 'us' is constructed as an in-group identity, the 'other' or the concept of 'them' is constructed as an out-group identity [313]. Therefore, polarisation and opposition are created by emphasising the differences between 'us' and 'them'. This may occur, for example, through the use of language to convey positive self-representation and negative representation of the 'other' as an out-group that is undesirable [336]. Although othering language may not contain explicitly hateful words, it does convey the desire to distance different groups [63], and within it there is an inherent promotion of discrimination and division of societal groups, fostering widespread societal tensions.

In machine learning research, the principle of othering has been identified by Burnap *et al.*[65] as a useful feature for classifying cyberhate based on religious beliefs, specifically for identifying anti-Muslim sentiment.

In their work, each tweet was computationally transformed into a list of all the individual words (tokens) in the tweet. Their interpretation for classifier improvement was based on the thinking that three-token terms (tri-grams) may represent ’othering’ and incitements to retributional action, such as ’send them home’ or ’get them out’. Applying a typed dependency parser to the previously mentioned examples (e.g. ’send them home’) can indicate relationships that can be classified as othering behaviour, e.g. (nsubj(home-5, them-2)). This essentially distances ’them’ from ’us’ through the relational action of removing ’them’ to their ’home’, as perceived by the author of the tweet. Indeed, they established the idea that othering language may be an important factor for improving a hateful classifier. However, their approach trained the classifier on the probability of the n-gram of each typed dependency occurring in a hateful or antagonistic tweet. Their effort involved interpreting certain statistically effective linguistic features but they did not test these features with machine classification algorithms and state-of-the-art features such as semantic features (that may capture similarities between hateful terms).

So far, this discussion has focused on novel methods and data to improve the detection of hate. Next, the discussion moves on to consider how to manage and disrupt the propagation of hate.

2.5 Online Hate Exposure and Spread Characterisation

Individuals and groups have increasingly used the internet to express their ideas, spread their beliefs, and recruit new members [193]. Thus, as online social media enables individuals and groups to spread ideologies and even advocate hate crime, it is essential to study the online structure, communication and connectivity of online communities in order to determine users' exposure to hateful ideologies that could influence their own views and actions.

The detection of hate online has been widely discussed from the perspective of content analysis. However, the study of hateful networks on social media has received limited attention in the literature. A study of such networks could be valuable in the context of concern about exposure to, and contagion of, online hateful and offensive narratives in social media. Several studies have applied Social Network Analysis (SNA) methods to the content of hateful networks communicating on the social media platform Twitter in order to use connectivity information as an indicator that a user is posting offensive content [273, 19]. Others have focused on SNA analysis of the retweets network to measure diffusion [278, 273]. However, there is yet to be a study of *multiple* hateful networks which aims to understand whether there is evidence of similar of 'levels of friendship', and therefore general exposure to the hate, nor to similar levels of propagation behaviour and therefore general contagion effect. On the Twitter platform, the hateful *followers' network* represents the user community directly exposed to hateful content [12]. This network is a subset of users who receive information directly from each other. Furthermore, the hateful *retweet network* is a construct formed by users who propagate cyberhate to their own followers [301], thereby passing on hateful narratives from the people they follow - a form of cyberhate contagion.

From a network analysis perspective, hateful content is produced by *nodes*, the nodes in the network are the people and groups while the links *edges* show relationships (e.g.

follow/friend) or flows between the nodes (e.g. retweets)[212].

For the characterisation of hateful networks, network metrics or 'network measures' [325, 169] are metrics that allow the quantification of a network as a whole and are helpful for comparing and classifying a set of networks. Also, they are useful when one network has undergone some topological changes (e.g. node/edge insertion or removal) [226]. For example, Giant Component (GC), density, degree distribution, clustering coefficient, reciprocity and diameter are among the numerous network measures. The size of the GC can reveal the maximum number of people who can be (directly or indirectly) reached by any other node in the same component, while the density is the ratio between the number of edges in the graph and the total number of possible edges[360]. The average shortest path is a direct measure of how information travels throughout the network, while the average degree and degree distribution focus on how information travels throughout the network [239]. The clustering coefficient measures how some of the nodes can form dense groups in which each element has strong connections with the others [353], while reciprocity reflects the measure of the likelihood that nodes in a directed network are mutually linked.

Also, there are many measures of node centrality in a graph to capture the importance of a node within such a structure. For example, the degree centrality, the betweenness centrality, and the eigenvalue centrality of a given node [324] or a group of nodes [112] are frequently used centrality measures. Degree centrality for individual nodes provides the number of direct links they are involved with and helps identify leaders which have the (almost) highest number of links within the network [226]. Betweenness centrality is a measure of accessibility - that is the number of times a node is crossed by shortest paths in the graph, which is useful for finding the individuals who influence the flow around a system [226], while eigenvector centrality is important as a connectivity measure for high information diffusion. Degree centrality measures the number of connections a node has but disregards the nodes to which these connections are established. Eigenvector centrality modifies this approach by giving a higher centrality score to those connections which are made with those nodes that are themselves central [210].

Table 2.6 presents a summary of previous hate-related network studies according to the platforms, network metrics or applied metrics, the goal of the study, and the number of examined networks.

Table 2.6: Summary of the studies that analyse the hateful networks in terms of hate diffusion, hate exposure and community detection

Study	**Platform**	**Applied metrices**	**The goal of the study**	**number of examined networks**
Gerstenfeld *et al.*[128]	Web	Link count	Hate exposure	1
Zhou et al[359]	Web forum	Link count	Hate exposure	6
Chau *et al.*[77]	Web forum	average shortest path length, clustering coefficient, giant component degree, betweenness, and closeness.	Hate exposure	28
Wiil *et al.*[327]	Web forum	Density, average shortest path length, clustering coefficient, degree distribution, degree, betweenness, eigenvalue, and closeness.	Hate exposure	1
Mathew *et al.*[213]	Gab	the average path length	Hate propagation	1

Table 2.6 Continued: Summary of the studies that analyse the hateful networks in terms of hate diffusion, hate exposure and community detection

De *et al.*[99]	Twitter	Frequency of interact within their own group, eigenvector, degree and betweenness	Hate propagation	1
Ting *et al.*[310]	Facebook	Average shortest path length, Diameter, density, degree centrality, closeness centrality, betweenness centrality, clustering coefficient	Communities detection	1
Wadhwa *et al.*[317]	Twitter	Time interval and cluster visualisation	Communities detection	1
Ribeiro *et al.*[272]	Twitter	Degree, betweenness and eigenvalue	Hate propagation	1
Xu *et al.*[343]	Web forum of complex systems	Average path length, average clustering, average Degree, degree distribution, link density, assortativity.	Hate propagation	4

Previous research on Twitter networks [317, 273] has been aimed mainly at studying only the propagation of the hate on one network. However, studying the propagation of cyberhate on multiple networks allows for a review of the differences and commonalities among those networks. The commonalities of hateful networks enhance our understanding of how to detect hateful networks and repress their ideologies. Analysing different networks that belong to different hateful events can provide information relevant for improving situational awareness and predicting the behaviour of the event in social networks, thereby providing greater support to the decision-making process [223]. Despite the fact that Chau *et al.* [77], Xu *et al.*[343], Wiil *et al.*[327] and Ting *et al.*[310] applied a wide range of metrics for the purpose of characterising their hateful networks, they focused their studies on web fora and Facebook, which are structurally different from Twitter. Among the many existing social networks, Twitter currently ranks as one of the leading platforms and is one of the most important data sources for researchers [173]. It is one of the top five social media platforms used in the UK [289]. Twitter is different from Facebook, Instagram, Snapchat and LinkedIn in that it is characterised by its short message limit (now 280 characters) and unfiltered feed which can be 'retweeted' (propagated) across a very open network. While Facebook is largely closed. Twitter's usage has quickly escalated, especially amid events, with an average of 500 million tweets posted per day. This feature has led Twitter to evolve into a popular tool for short and immediate commentary on real-time happenings, including both personal and news events [171].

Another communication feature of Twitter is the hashtag: a metatag beginning with # that is designed to help others find a post, which emphasises the importance of widely communicating information on Twitter [109]. In addition, such a feature provides data filtering, e.g. retrieving tweets in a specific language or from a certain location. This flexibility in retrieving data encourages developers to perform research and analysis using Twitter. Additionally, Twitter data is structured in such a way that all information regarding a tweet is rolled into one block using e.g. the CSV, Json, etc file format. A block consists of many fields relating to user information, tweet description and re-tweet status. This type of structure eases difficulties in mining for specific information such as tweet content while ignoring other details such as user or re-tweet status.

Furthermore, corpora constructed from social media and websites other than Twitter are rare, making it difficult for analysis of hate speech to cover the entire landscape [118]. At the onset of this research project, Twitter was providing its data via a number of Application Programming Interfaces (API). In contrast, the aftermath of the **Cambridge analytica 'data breach'** has led to certain social media platforms limiting data provided through their Application Programming Interfaces [17]. In recent years, social networks (and especially Twitter) have been used to spread hate messages. In addition, unlike many of the other platforms, such as news feeds, blogs and emails, it has an explicit social (i.e. subscription) network [277]. From a researcher perspective, Twitter is a logical source of data for such analysis given that users of social media are more likely to express emotional content due to deindividuation (anonymity, lack of self-awareness in groups, disinhibition) [115] and, therefore, form hateful communities that propagate their ideologies.

Unfortunately, Twitter has been heavily abused by some who make it their mission to demonise and preach hatred against other religions [328]. On Twitter, an important event, e.g. a hateful event, can be expected to trigger more informational tweeting[156]. In support of this, Twitter commentaries have been shown sometimes to quite closely reflect offline events, such as political deliberations and religious events[311]. Also, Williams *et al.* [329] mentioned that there is an increase in online anti-Muslim and anti-Black speech on Twitter associated with an increase in racially and religiously aggravated violence, criminal damage, and harassment, showing that Twitter is now part of the formula of hate crime. It is therefore reasonable to conduct an information-flow analysis of Tweets posted by users.

2.5.1 Cyberhate Management by Twitter

Twitter recognises that if people experience abuse on Twitter, it can jeopardise their ability to express themselves. It is committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalised. For this reason, Twitter prohibits behaviour that targets individuals or groups with abuse based on their perceived membership in a protected category. They review and take action against reports of accounts targeting an individual or group of people with any of the following behaviour, whether within Tweets or Direct Messages:

- Prohibit content that makes violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., 'I will kill you'.

- Prohibit content that wishes, hopes, promotes, incites, or expresses a desire for death, serious bodily harm, or serious disease against an entire protected category and/or individuals who may be members of that category.
- References to mass murder, violent events, or specific means of violence where protected groups have been the primary targets or victims.
- Prohibit inciting behaviour that targets individuals or groups of people belonging to protected categories. This includes content intended to (i) incite fear or spread fearful stereotypes about a protected category, (ii) incite others to harass members of a protected category and (iii) incite others to discriminate in the form of denial of support to the economic enterprise of an individual or group.
- Repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone.
- Twitter considers hateful imagery to include logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, gender identity or ethnicity/national origin.

Twitter takes action against behaviour that targets individuals or an entire protected category with hateful conduct, as described above. Targeting can happen in a number of ways. When determining the penalty for violating this policy, Twitter considers a number of factors including, but not limited to, the severity of the violation and an individual's previous record of rule violations. The following is a list of potential enforcement options for content that violates this policy:

- Down-ranking tweets in replies, except when the user follows the Tweet author.

- Making tweets ineligible for amplification in top search results and/or on timelines for users who don't follow the tweet's author.
- Excluding tweets and/or accounts in email or in-product recommendations.
- Requiring tweet removal, e.g. asking someone to remove the violating content and serving a period of time in read-only mode before they can tweet again.
- Suspending accounts whose primary use it has determined is to engage in hateful conduct as defined in this policy, or who have shared violent threats.

Social network analysis, particularly on the Twitter platform, has been applied in a range of studies, e.g. student interaction [294], quantifying influence on Twitter [37], political community structure and emotions [81, 146].

In terms of online hate on the web forum, Gerstenfeld *et al.* [128] analysed 157 extremist sites and found links between most of these websites, and Zhou *et al.* [359] also investigated web communication and analysed the content and links of hate groups. In their research, they found that the main objective of these websites is to spread and promote ideas, such as those of white supremacists and neo-nazis. Moreover, Chau *et al.* [77] used SNA techniques to analyse hate groups on the internet, formulating hypotheses around the specific features of each site. They showed that the network of bloggers in hate groups is decentralised. Also, they found that the number of 'hate' bloggers has increased steadily over a number of years.

Recently Mathew *et al.*[213] introduced a study that looked into the diffusion dynamics of posts made by hateful and non-hateful users on Gab. They collected a large dataset of 341K users with 21M posts and investigated the diffusion of the posts generated by hateful and non-hateful users. They observed that the content generated by the hateful users tended to spread faster, farther and reach a much wider audience when compared to the content generated by non-hateful users. Also, an important finding was that hateful users were far more densely connected among themselves compared to non-hateful users. However, unlike Twitter, Gab promotes 'free speech' and allows users to post content that may be hateful in nature without any fear of repercussion. Thus, this network is not comparable with a Twitter network as the hateful content would spread on Twitter with restrictions.

On Twitter, previous research has been aimed mainly at detecting hateful, offensive, abusive and aggressive speech on the platform using information about network activity. Such studies have analysed user network activity on Twitter to detect cyberhate by considering specific attributes of online activity using machine learning classifiers. An example is Chatzakou *et al.* [75] who detected Twitter aggressors and bullies automatically; equally, Ribeiro *et al.* [273] detected hateful users and Ting *et al.*[310] focused on hate group detection. Burnap *et al.* [64] specifically looked at retweet virality following a terror attack - a likely trigger event for hateful responses - and found that sentiments expressed in tweets were statistically significantly predictive of both size and survival of information flows of this nature. Wadhwa *et al.* [317] aimed to uncover/identify hidden radical groups in online social networks, providing evidence of the ability to discover subgroups. Ribeiro *et al.*[272, 273] aimed to define a user-centric view of hate speech by examining the difference between user activity patterns and network centrality measurements in the sampled graph. They discovered that hateful users were more central in the retweets network, and therefore identifiable as key influencers within the network.

2.6 Online Hate Propagation and Disruption

Cyberhate is disseminated by both groups of haters (hateful communities or networks) and by those acting independently. The main objectives of such groups and individuals are to recruit and link like-minded people in support of their cause and spread hateful content [46]. Exposure to cyberhate leads to deteriorating intergroup relations [348]. Previous studies have shown that exposure to online hate content is associated with seriously violent behaviour[347, 245]. To overcome the problem of hate speech exposure and the spread of hate speech, we need to observe and manage the content disseminated by such groups. This chapter now goes on to explore the literature related to limiting such communities from propagating their content by removing the most effective nodes that underpin its propagation.

Network node removal is a well-known technique for destabilisation of networks[69]. From a topological perspective, node removal is always more effective in atomising complex networks causing more damage per elimination than edge removal, since the deletion of a single node from the network results in the elimination of all the links attached to it [158, 90].

Classic results focusing on the problem of node removal indicate that many real networks show a 'robust yet fragile' nature, i.e., they are robust to random node removal but very fragile to attack of the nodes with key connectivity roles in the network. *Boldi et.al.* [53] found that there is a clear structural difference between social networks (such as Twitter) and web graphs, and therefore it is important to test node removal strategies until a significant fraction of the nodes has been removed.

2.6.1 Single Node Removal Strategies

Yip *et al.* [351] examined the structural properties of the networks of personal interactions between cybercriminals in carding forums. They found that carding social networks are not scale-free[2], as the degree distributions are log-normal[3], which has important implications for network disruption. It is widely accepted that scale-free networks are particularly resilient to random node removals, but highly vulnerable to targeted attacks; this is due to there being only a small fraction of nodes possessing the

[2] A scale-free network is a network indicating that the vast majority of nodes have very few connections, while a few important nodes (called Hubs) have a huge number of connections.

[3] log-normal (or lognormal) distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed (few nodes of very large degrees and a great many nodes of small degrees).

majority of links. In their study, they did not use any node removal strategy; instead, they used the implication of the degree distributions characteristics. Petersen *et al.* [264] examined removing the highest degree nodes based on the distance between the entire network's nodes. Their work proposed a node removal algorithm for a criminal network. As part of their study, they found that removing the high degree nodes had an impact on enlarging the distance between the criminals. Wiil *et al.* [327] introduced a study that analysed the importance of links in terrorist networks. This study showed that removing nodes destabilised the network, noting that both the importance of nodes and links should be considered. All this previous work has been examined on web fora, which are structurally different to that of the Twitter platform [170]. For the Twitter platform, a node removal strategy was applied on political networks and showed that SNA metrics could be used to evidence the impact on the network connectivity [168]. An attempt by Xu *et al.* [343] found that terrorist networks on the web were more vulnerable to attack on the bridges that connect different communities, than to attacks on their hubs. They applied two removal strategies on websites' networks: a hub-based and a bridge-based strategy.

Deciding which node or group of nodes to remove depends upon what information is available on the importance of the nodes in the network. Removing nodes to reduce a network's connectivity (and, therefore, stem the flow of content) is something that has been widely addressed in previous research. For example, these strategies are used for breaking complex networks [95], the spread of computer viruses [241], and spam prevention [87]. Two aspects to consider when applying node removal strategies include: (i) which node removal strategy to apply, and (ii) the impact measure of the strategy on the network's structural properties. When the aim is to break down a network, it makes much more sense to target certain nodes instead of removing them randomly. This is the case, for example, when battling against a virus or attempting to dismantle a criminal network.

Node removal includes several strategies. Table 2.7 summarises the studies that applied different node removal strategies to different types of networks. Here, for example, a widely used strategy is degree centrality. Several measures have been used to evaluate the impact of node removal on the robustness of networks [50]. For example, a strategy can be measured according to how the diameter of the network changes, or how the largest components' sizes change over periods of removal [242]. Table 2.7 also shows the studies that have adopted two widely used indicators, i.e., Giant Component (GC) and network distance. Identifying and removing important nodes from networks are great challenges in many real-world applications. For an example, in science co-authorship networks, determining which node removals produce higher information spreading reduction may help us to identify the nodes/scientists who are making the greatest contribution to knowledge and idea spreading [16, 182, 255]. Furthermore, to destroy criminal networks, it is desirable to break a network into smaller components or increase the distance between node s[94].

These findings may furnish useful tools for designing policies facilitating the activities of these haters, who act as 'influential spreaders' in the network. These analyses can be useful for finding which criminals play a major role in shaping information delivery within criminal networks, thus providing knowledge for investigative policies [15, 94].

Studies which specifically aim to reduce connectivity in hateful networks by removing the nodes are rare, and, as of yet, no study exists that examines such online social networks. Such a study is needed to understand the critical nodes that have an effect on the spread of cyberhate and therefore could be removed from a network to reduce the flow of hateful information, and subsequently reduce the harm caused by hateful communities on Twitter. Intervention methods could include the possibility of identifying contagion pathways in hateful networks and evaluating the reduction in exposure of the network's users to receiving hateful content, in the same way that we might expect the spread of a traditional offline virus to be contained.

Table 2.7: Summary of the studies that use different network prevention strategies

Study	**Platform**	**Type of network**	**Applied strategy**	**Test the impact on**
Holme ***et al.***[150]	Scientific collaborations and Internet traffic	Social Network	Degrees and betweenness centralities	Giant Component (GC) and distance
Jahanpour ***et al.***[160]	airplane hijackers' network	Complex Network	Degrees, eigenvalue and betweenness centralities	Giant Component (GC), the shortest paths, network reciprocal distance, Average node coverage, clustering coefficient and shortest distance homogeneity
Williams ***et al.***[335]	Killer whale	Social Network	Highest degree	Giant Component (GC)
	Vaccine Network	Epidemics Network	Out degree centrality	Giant Component (GC)
Xu ***et al.***[343]	Web forum	criminal networks	hub removal, bridge removal	path distance

Table 2.7 Continued: Summary of the studies that use different network prevention strategies

Petersen *et al.*[264]	Web forum	Criminal Networks	degree centrality	path distance
Boldi *et al.*[53]	web graphs	-	Random, largest-degree, root nodes, pagerank, label propagation and betweenness/harmonic centrality	path distribution
Mourier et al[231]	Blacktip reef shark	Social Network	Highest degree	Giant Component (GC)
Colladon *et al.*[87]	Business emails and Twitter	Spam	Closeness, degree, betweenness centrality, average response times, activity, contribution index and nudges	Average distance, clustering coefficient, average degree.
Jurgens *et al.*[168]	Twitter	political communication	Highest degree	Centrality entropy(degree of reachability)
Newman *et al.*[241]	Email network	Network Viruses	Highest degree	Outbreak size

Table 2.7 Continued: Summary of the studies that use different network prevention strategies

Albert *et al.*[23]	Web forum	metabolic network	Average degree	Average shortest path
Wiil *et al.*[327]	web links	terrorist network	Higher link betweenness	Efficiency (sum of the shortest paths connecting each pair of nodes is computed.)

2.6.2 Hybrid Node Removal Strategies

Several research studies have applied a combination of two or more node removal strategies - ’hybrid’ - in order to establish the importance of the nodes in the network. Bellingeri *et. al.* [40] combined first-degree and second-neighbour-degree strategies to find the crucial nodes in real networks. They also examined degree-based and betweenness-based strategies. They found that the betweenness strategy alone was the most effective strategy for reducing the size of the biggest component. Additionally, Wang et al. [318] asserted that the importance of nodes is related to the degree of nodes and their neighbours, and proposed a new algorithm to rank the importance of network nodes. Ruan et al. [349] proposed a combined node importance ranking algorithm which only requires the centrality of the nodes to be obtained and combined with the neighbourhood information within two hops of the node. The algorithm demonstrated that the bigger the degree of a node and the fewer connections between neighbouring nodes, the more important the target node. Indeed, it would be interesting to examine a combination of two or more (’hybrid’) node removal strategies for hateful Twitter networks, whether in relation to followers of networks or retweeters of networks.

2.6.3 Removal Strategies on the Bipartite Networks

Before 2010, researchers mainly studied intervention methods and the robustness of generic networks or (one-mode networks) [68]. A social network analysis, particularly in relation to intervention strategies, can also be viewed and applied according to a user's graph and their affiliation to other people. In order to capture more information on the stability of complex systems, scientists suggest multilayer (the simple form is called multipartite) network modelling [49, 98], e.g. two-mode networks or bipartite networks [238]. A bipartite graph is a graph with two sets of vertices which are connected to each other, but not within themselves. More formally a bipartite graph G = (V_1, V_2, E) consists of a set of vertices V_2 a disjoint set of vertices V_1 and a set of edges E $\in$ V_1 X V_2. For a generic network ('a one-mode network'), all the nodes follow the same degree distribution. The two-mode network has two node sets. Therefore, the nodes in each node set do not need to follow the same degree distribution. This means that removal strategies that are applied to a generic network ('one-mode network') or bipartite networks ('two-mode network') have different impacts on a network's connectivity. Several studies have applied intervention strategies to bipartite networks [346, 288, 258]; however, this field has attracted limited attention. Xuan *et. al.*[344] studied the robustness of bipartite task-oriented social networks, in the absence of workers (attack). They proposed four attack strategies, including efficiency-based attacks, centrality-based attacks, diversity-based attacks and influence-based attacks. They found that their bipartite networks were robust in the case of a centrality-based attack. As no study has yet applied the node removal strategies to a bipartite hateful network, it would be interesting to examine the node removal strategies in relation to bipartite versions of hateful networks.

2.7 Research Gap Analysis

So far in this chapter, we have reviewed the literature on three related concepts: contextual cyberhate classification; characterising hateful networks; and disrupting hateful networks. Here we will summarise the key gaps identified for each concept.

2.7.1 Hate speech classification

There have been several attempts in the area of cyberhate classification to automatically identify and quantify cyberhate by using different approaches, such as lexicons, syntactic and semantic features. For lexicons [131, 321, 105, 287], the problem which persists with existing approaches is that they depend on the existence of the co-occurrence of direct hate features to decide the polarity of a specific tweet, which might mean that indirect/implicit hate speech cannot be identified. Additionally, they suffer from a high false-positive rate due to the model which classifies sentences. Syntactic methods include [136] such as BoW [63, 136, 187], n-grams[36, 63, 96, 136, 243, 323], TF-IDF[104, 36], Part-of-speech (POS)[104, 136, 62], Latent Dirichlet Allocation (LDA)[342, 12],Typed Dependencies (TD)[179, 153]; The works which use these methods have not considered the detection of implicit hate speech. Semantic learning, such as text embedding, aims to train a model that can automatically transform a sentence/word into a vector that encodes its semantic meaning. It is an opportunistic feature which enables the meaning of the text to be extracted. However, studies which use this feature [183, 36, 106, 243] for hate speech classification have not yet tackled the classification of text that does not contain clear hateful words and this would have an impact on classification accuracy (e.g. *get them out of our country*). So, there is a gap in terms of utilising previous approaches in order to detect indirect or implicit online hate. Indeed, examples of implicit hate require an additional layer of qualitative context that sits above combinations of individual words. Thus, there is a need for a model capable of understanding and detecting implicit hate, perhaps by focusing on how language conveys hateful or antagonistic sentiments. For example, Burnap et al. [63] suggested that utilising psychological theories in order to detect cyberhate would contribute to improving the detection process. They used text parsing to extract typed dependencies, which represent syntactic and grammatical relationships between words, and are shown to capture 'othering' language. They showed that using typed dependency features is consistently improving machine classification for different types of cyber hate beyond the use of a Bag of Words and known hateful terms. However, their work focused on post-classification to interpret some of the statistically useful linguistic features. It is yet to be used as a theoretical foundation for feature engineering and tested with machine classification algorithms and state-of-the-art features such as semantic features (that may capture the similarity between the hateful terms). As 'othering' language is a type of implicit hate speech, we could consider othering language, and its stereotype, as an extra feature layer that can capture a type of implicit hate. Our hypothesis here is that: *linguistic features associated with othering language provide an additional set of qualitative features that will improve classification performance.* This leads to the first research question:

RQ1: To what extent can using othering and ITT theories along with embedding learning drive the development of new features for classifying cyberhate and improve the performance of machine learning for cyberhate detection?

This research builds on previous studies by developing new evidence that the complex and nuanced 'us and them' narrative emerging on social media can be captured to improve cyberhate classification. A suggested solution is to use a syntactic feature such as TD and POS in order to extract othering language and its patterns. Then, paragraph embeddings would infer semantic similarity between features to create a model that represents 'othering' language, which is used for the purposes of cyberhate classification. Djuric *et al.* [106] presented the paragraph2vec approach to classify language in users' comments. They demonstrated that using the vector space features, particularly sentence (paragraph) embedding, improved the performance of the classifier compared to not applying vector space features. Our expectation is that if samples that contain ’othering’ language or its patterns (e.g. verb-pronoun combinations) in a hateful or antagonistic context are aligned in similar feature ’spaces’, this would increase the probability of the machine classification method labelling any new samples which exhibit these features as cyberhate.

2.7.2 Hateful networks’ characterisation:

Dealing with cyberhate is not limited to detecting or classifying the hateful context, but it is also related to those users and their groups who post this content. Indeed, there has been limited attention given to characterising these groups and how they propagate their content. In fact, some previous research, on Twitter networks [317, 273] was primarily aimed at studying only the propagation of hate on one network. However, there is a gap in this approach because studying the propagation of cyberhate on multiple networks would allow for a review of the differences and commonalities among those networks. Moreover, there is another gap in that there has been very little attempt to study the networks in terms of propagation (retweet network) and hate speech exposure (follower networks). It is important to note that people who are exposed to hateful content will not necessarily spread hate. In addition, it seems there is yet to be a study of *multiple* hateful networks with the aim of understanding whether there is evidence of similar ’levels of friendship’, and therefore a broad exposure to hate, or similar levels of propagation behaviour and therefore a general contagion effect. Moreover, evaluating the results of network characterisation may benefit from comparison with another risky network, something which has not been considered in previous studies. Our expectation of characterising multiple hateful networks is that we may find some similarities among the hateful networks regarding the friendship exposure level in the followers’ networks and propagation level in terms of retweet networks. Therefore, we hypothesise that:*hateful networks are similar in terms of online hate exposure and online hate propagation, and more connected compared to another ’risky’ networks.*

The above limitations and hypotheses led to the formulation of the second and third research questions of this thesis:

RQ2: By studying multiple hateful networks on Twitter, is there evidence of similar of 'levels of friendship' across multiple hateful networks, and therefore a general measure of exposure to the cyberhate?

RQ3: By studying multiple hateful networks on Twitter, is there evidence of similar levels of propagation behaviour and therefore a general contagion effect?

Note that a network is labelled as a hateful network if all users belonging to that network have posted tweets that human annotators agree should be classified as containing evidence of hateful content.

2.7.3 Hateful networks' disruption:

Classifying hateful content and characterising hateful networks requires an additional reaction which involves disrupting these networks so their content propagation is disrupted earlier. There are several works which focus on the disruption of criminal networks, such as those by Yip *et al.* [351], Da *et al.*[94] and Petersen *et al.* [264]. Additionally, Wiil *et al.* [327] and Xu *et al.* [343] introduced a study that analysed the importance of links in terrorist networks. They found that terrorist networks on the web were more vulnerable to attacks on the bridges that connect different communities than to attacks on their hubs. The majority of previous studies have focused on web networks other than Twitter. Our concern here is that the node-removal strategies may behave differently for different network topologies. It has been demonstrated by Boldi *et al.* [53] that there is a clear structural difference between social networks (such as Twitter) and web graphs, and therefore it is important to test node removal strategies until a significant fraction of the node has been removed. So, the limited attention paid to the disruption of hateful networks on Twitter is an additional gap, though one study was conducted by Da *et al.* [95] which was implemented on the Twitter platform to research the propagation of hate; however, there is also a need to investigate how to disrupt exposure to hate among hateful users. Another gap is that researchers generally examine one network while there is, in fact, a need to examine more than one network to generalise the results. It is also important to compare and contrast the results of disrupting networks with another 'risky' network in order to evaluate the results. Previous studies have yet to propose disruption methods, specifically removal strategies for network nodes, to prevent cyberhate exposure and spread based on the examination of multiple networks from Twitter [31, 29, 230]. Therefore, we hypothesis that: *applying node removal strategies (disruption strategies), depending on the node role in the network, will reduce network connectivity (exposure reduction) and diffuse the spread of hate (contagion reduction).* This prompted the following question:

RQ4: According to the structural characteristics of networks, which node removal strategies would be most effective in decreasing the propagation of hateful content?

Additionally, no study has yet been undertaken that examines combined or 'hybrid' node removal strategies for hateful Twitter networks, whether among followers of networks or retweeters of networks, prompting the following question:

RQ5: According to the structural characteristics of networks, is a combination of two (hybrid) node removal strategies more effective in decreasing the propagation of hateful content compared to applying only a single node removal strategy?

As part of examining node removal strategies on hateful networks, we need to examine the node removal strategies for multilayer (the simple form is referred to as bipartite) hateful network modelling. Since no study has yet applied node removal strategies to a bipartite hateful network, the sixth research question in this thesis is as follows:

RQ6: Does applying the node removal strategies to a bipartite version of hateful networks improve the node removal strategies in terms of detecting the most important users?

2.8 Conclusion

This chapter has explored the background knowledge relating to detecting and countering implicit and explicit online hate speech, or cyberhate. To gain a full understanding of the topic, the existing work related to defining online hate was reviewed, followed by clarification of what is meant by online hate speech in this thesis. Recent techniques used for detecting online hateful content from a feature extraction perspective and also from a machine classification perspective were reviewed. This was followed by paying particular attention to hateful networks' characterisations. In addition, this chapter concluded by providing an exploration of studies which have investigated content propagation on hateful networks. Six research questions were identified, Chapter 4 contributes towards addressing **RQ 1:** by developing the first contribution **C1**. Chapter 5 contributes towards addressing **RQ 2** and **RQ 3** by presenting an extensive analysis of hateful networks on Twitter and resulted in the second **C2** and the third **C3** contributions. Chapter 6 contributes towards addressing **RQ 4,5 and 6** by introducing the fourth contribution **C4**. Finally, This chapter highlights the major knowledge gaps for each area.

The next chapter provides a general overview of the approach adopted in this thesis and a description of the dataset collected is also given.

Chapter 3

Research Design

3.1 Introduction

This chapter provides an overview of the proposed methods for hateful content classification, hateful network characterisation and hateful network disruption.

The research methodology is presented in Section 3.2, and the research gaps analysis is in Section 3.3. The proposed framework is presented in Section 3.4. The data used are described in Section 3.5. Finally, a summary of the chapter is provided in Section 3.6.

3.2 Research Methodology

Research methodology is defined as a systematic way to solve a research problem by collecting data using various techniques, providing an interpretation of the collected data, and drawing conclusions about the research data. A research method is fundamentally the blueprint of the research or study. Generally there are three kinds of approaches or research methods namely qualitative, quantitative and mixed. These methods are used to gather data and resolve issues that emerge during the process of data gathering [320]. An example of the qualitative method is Design science research in which the object of study is the design process. Quantitative methodologies include experiments, observation and structured interviews.

In this research, to verify the hypothesis, we applied the Design Science Research Methodology (DSRM) introduced by Peffers et al. [265] as depicted in Figure 3.1. Each step is described and related to these PhD thesis chapters as follows:

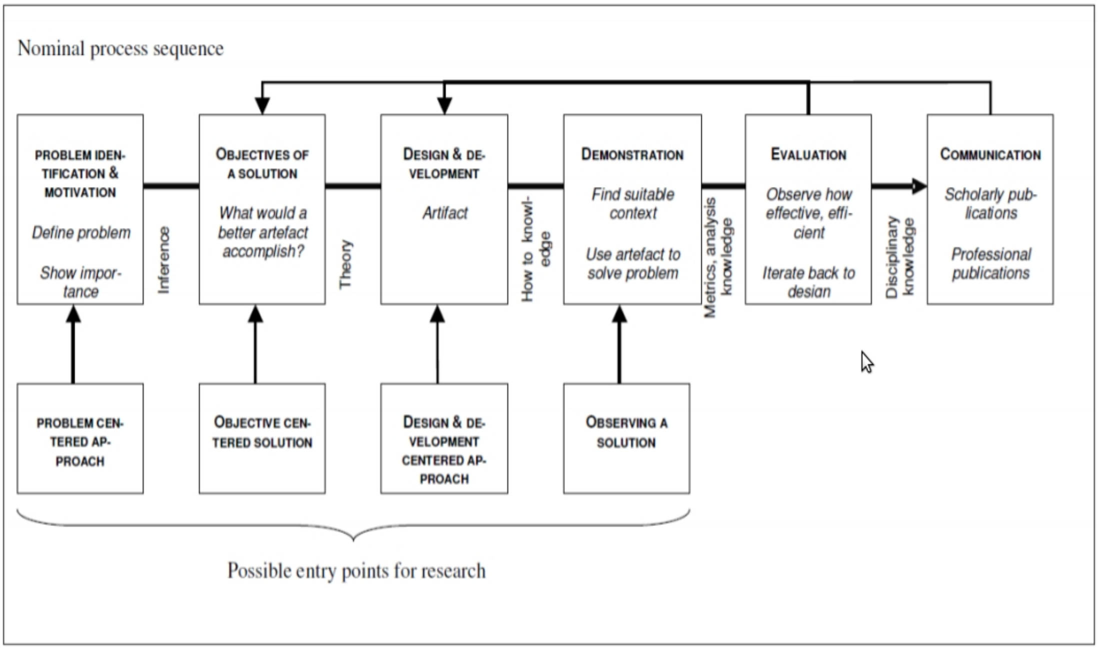

Nominal process sequence
PROBLEM IDENTIFICATION & MOTIVATION
Define problem
Show importance
Inference
OBJECTIVES OF A SOLUTION
What would a better artefact accomplish?
Theory
DESIGN & DEVELOPMENT
Artifact
How to knowledge
DEMONSTRATION
Find suitable context
Use artefact to solve problem
Metrics, analysis knowledge
EVALUATION
Observe how effective, efficient
Iterate back to design
Disciplinary knowledge
COMMUNICATION
Scholarly publications
Professional publications
PROBLEM CENTERED APPROACH
OBJECTIVE CENTERED SOLUTION
DESIGN & DEVELOPMENT CENTERED APPROACH
OBSERVING A SOLUTION
Possible entry points for research

Figure 3.1: Design Science Research Methadology (source: Peffers el al. [265])

- **Problem identification and motivation**: This phase involves a critical and deep understanding of hate speech classification, hateful networks characterisation and disruption - and the related research areas. The first step of this phase involves the identification of gaps in the related literature, as presented in Chapter Two. In the second step, the research hypothesis statement and the research questions are identified as presented in Chapter One. The third step requires the choice of data source and the development tools used to test the hypotheses. The last step involves a time plan for the research by dividing the main problem into tasks and identifying the required milestones.

- **Objectives of the solution:** In this step, the problem definition in the previous step is used to propose the objectives of the solution. For hate speech classification, this research aims to develop a classification framework that takes advantage of the existence of othering language in a hateful tweet to identify implicit and explicit hate speech. For hateful networks characterisation, this research aims to understand the characteristics of hateful online social networks in order to help to understand individuals' exposure to hate and cyberhate propagation. Moreover, in terms of hateful network disruption, this research aims to reduce online hate exposure and propagation.

- **Design and development:** This step aims to design a solution to the problem and develop it. In this research, The 'othering feature set' was proposed to develop a classification framework that takes advantage of the existence of othering language in a hateful tweet. In addition, multiple hateful networks were collected and built in order to understand the characteristics of hateful online social networks in a small non-representative sample in order to help understand individuals' exposure to hate and cyberhate propagation. Also, node removal strategies were applied to reduce online hate exposure and propagation. This step has been explained in detail in Chapters Three, Four, Five and Six. The entire design of the general framework is explained in Chapter Three.

- **Demonstration:** This step involves using the developed framework in a proper context. In this thesis, different experiments have been carried out in Chapters Four, Five and Six using samples of Twitter datasets to demonstrate the effectiveness of the proposed framework. The Twitter data are made up of contextual tweets that have been posted by hateful users and the users' profile information needed for building hateful networks, such as a followers list. Chapter Four introduces a novel feature set for cyberhate classification based on the use of two-sided pronouns and the use of pronoun patterns such as verb-pronoun combinations. A wide range of classification methods were implemented to compare our novel approach that fuses embedding learning with an 'othering' narrative, to state-of-the-art methods. The chapter also implements a qualitative analysis to demonstrate how the novel feature set can predict the vector space similarity between different kinds of hateful content. For Chapter Five, several hateful networks are characterised extensively, from the exposure to cyberhate (in follower networks) perspective to the propagation of cyberhate (in retweet networks) perspective. A range of network analysis metrics are used to compare and contrast baseline measures of connectivity and propagation across multiple hateful networks. Chapter Six develops strategies to identify nodes within hateful networks (user accounts) whose removal is empirically shown to reduce connectivity (largest component, density and average shortest path) in both the follower and retweet networks. Thirteen node-removal strategies, including random-based strategy, based on network connectivity, were tested on three network metrics: giant component size, density and the average shortest path. These strategies were applied to generic networks and bipartite networks.

- **Evaluation:** This step involves assessing the effectiveness of the method proposed compared to other methods. In this research, the evaluation experiments aim to measure the impact of the proposed solutions. For Chapter Four, the evaluation experiments aim to measure if the linguistic features associated with othering language provide an additional set of qualitative features that will improve performance in terms of classification. A wide range of state of the art classifiers were trained and evaluated using ten-cross validation, and the best performing classifiers were tested on an unseen dataset. The effectiveness of the classifiers is measured using recall, precision and F1. For Chapter Five, the evaluation revolved around comparing and contrasting different hateful networks characteristics and comparing a 'risky' network - characterised by language related to suicidal ideation. The evaluation process here examines if hateful networks are similar in terms of online hate exposure and online hate propagation, and more connected compared to another 'risky' network.

  For Chapter Six, the evaluation process aims to identify if applying node removal strategies (disruption strategies), depending on the node role in the network, will reduce network connectivity (exposure reduction) and diffuse the spread of hate (contagion reduction). A comparison between the impact of thirteen node removal strategies was also used to evaluate each strategy's performance. In addition, this was compared with a 'risky' network - characterised by language related to suicidal ideation.

- **Communication:** In this final step, researchers publish their contributions to the audience to elicit their feedback and convey the importance of the problem and its novelty. This thesis resulted in two publications: two journal papers. The publications are listed in the list of publications section.

3.3 General Framework

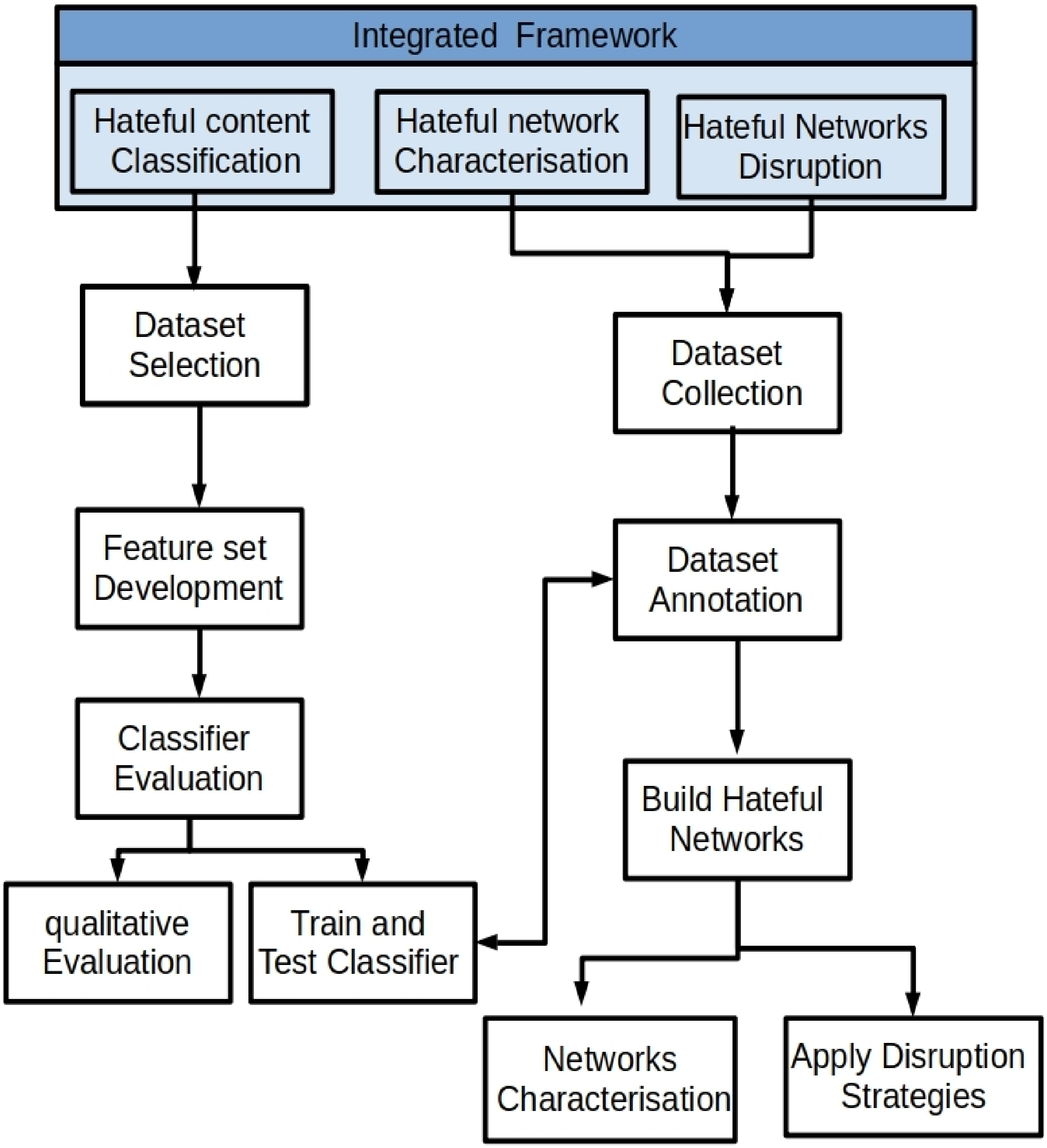

Integrated Framework
Hateful content Classification
Hateful network Characterisation
Hateful Networks Disruption
Dataset Selection
Feature set Development
Classifier Evaluation
qualitative Evaluation
Train and Test Classifier
Dataset Collection
Dataset Annotation
Build Hateful Networks
Networks Characterisation
Apply Disruption Strategies

Figure 3.2: The General Framework

This thesis introduces a new general framework for detecting online harms, as well as characterising and disrupting cyberhate networks. The general framework is divided into three sub-frameworks: cyberhate classification, cyberhate characterisation and cyberhate disruption. An outline of the framework is shown in Figure 3.2, and the different stages are described in more detail in Chapters Four, Five, and Six.

3.4 Data

3.4.1 Hateful Content Classification

In the context of hate speech automatic detection in this study, an instance is a Twitter post (Tweet) with a classification label. Regarding the number of instances per dataset, an appropriate sample size could have a wide range of magnitudes in existing literature. The majority of papers use between 1,000 and 10,000 instances [118]. For this study, two datasets were used for experimentation. To develop the othering feature set the dataset provided by Davidson *et al.* [96] was used. They collected tweets containing different types of hate and used crowdsourcing to further divide the sample into three categories: those containing hate speech, those with only offensive language, and those with neither. Annotators were asked to think not just about the words appearing in a given tweet, but also about the context in which they were used. They were instructed that the presence of a particular word, however offensive, did not necessarily indicate a tweet was hate speech. Each tweet was coded by three or more annotators. The inter coder-agreement score was 92%. They used the majority decision for assigning a label to each tweet. This resulted in a sample of 24,802 labelled tweets, with 21% of the tweets being coded as hate speech and offensive language (5% hate speech and 16% as offensive language) by the majority of annotators. This is referred to as our *Training Dataset* which contains 3161 non-malicious[1] samples and 5323 hateful samples - which is at the higher end of sample-size in previously published literature which used embeddings to classify cyberhate (e.g [101] 540 hateful samples, [125] 1037 hateful samples, [165] 712 hateful samples, [356] 413 hateful samples, [266] 1943 sexism /3166 racism samples).

To compare this work to the state-of-the-art in cyberhate classification, a second dataset was used for testing purposes, this was also done in previous work, see for example

[1] The term 'malicious tweets' refers to tweets that include any of the inappropriate behaviours outlined in Chapter 2 with the intent to hurt others.

[63]. The dataset was collected by [63] from Twitter and the method of data collection was random to ensure that the produced dataset was free from bias. The dataset was annotated using the CrowdFlower human intelligence task service with a single question: 'Is this text antagonistic or hateful based on a protected characteristic?'. The dataset comprises cyberhate directed at four different protected characteristics, as follows: sexual orientation - 1803 tweets, with 183 instances of offensive or antagonistic content (10.15% of the annotated sample); race 1876 tweets, with 73 instances of offensive or antagonistic content (3.73% of the annotated sample); disability 1914 tweets, with 51 instances of offensive or antagonistic content (2.66% of the annotated sample); and religion 1901 tweets, with 222 instances of offensive or antagonistic content (11.68% of the annotated sample). The authors conducted all of the necessary tests so as to ensure agreement between annotators for the gold-standard samples [63]. The amount of abusive or hateful instances is small relative to the size of the sample. However, these are random instances of the full datasets for each event and they are considered representative of the overall levels of cyberhate within the corpus of tweets [62]. The relative improvement in classification performance was evaluated using this dataset, which is referred to as our *Testing Dataset*.

3.5 Hateful Networks Characterisation and Disruption

When conducting a Social Network Analysis, relational data that reveals some kind of connection between the individuals or groups in the network is needed.

While Chapter 4 focused on the content of the tweets, this chapter studies the social network connectivity among users who posted the hateful tweet, and likewise on their communication. An example of the network connectivity is the user's followers and friends, whereas an example of the network communication is retweeting or mentioning a tweet. For author h, the followers are people who follow the author h. For example, $h \xleftarrow{follow} x$, means that x is a **follower** of h. There are several ways of communicating on Twitter including: 1) Retweets: sharing another user's tweet with your followers; 2) @Replies: directly replying to other user's tweets, which can grow into a conversation and 3) Mentions: referencing a specific Twitter user's username at any point in your tweet [299]. In this study, the interest lies in the followers and retweet action. Note that the retweet action is not necessarily performed by a follower; it could be performed by any users exploring Twitter (unless the tweet is protected or the author of the tweet blocks the user who want to retweet).

In order to collect and analyse hateful connectivity and communication posted to Twitter, accounts that were demonstrably posting hateful tweets need to be identified. Application Programming Interface (APIs) are the protocols that enable platforms, such as Twitter to give access to a controlled set of data. Twitter's API permits anyone with a Twitter account limited access to the recent retweets and followers of any public account.

These chapters uses data from anti-religious content. According to Prevent Strategy report by the UK government [9] and Heath *et al.* [145], anti-religious hate might have a connection with terrorism, extremism, radicalisation, etc. This suggests that understanding exposure to, and propagation of, anti-religious hate online may widen the researchers' horizons to comprehend groups who pose a risk to societal security.

These chapters uses data from two types of anti-religious content: Anti-Muslim and Anti-Semitic. These datasets contain religious content but are different from the previous chapter's dataset. The previous chapter's dataset did not contain users' profiles information (mainly users' names) needed for these chapters study to retrieve the links between the users (e.g.follow, retweet).

For the Anti-Muslim datasets, data was collected from Twitter that centred around two 'trigger' incidents. The first, which were collected by Burnap *et al.* [62], was related to the murder of Lee Rigby, a solider based in Woolwich, London, Data collection lasted two weeks following the terrorist attack committed on May 23rd, 2013; this data set was named 'Anti-Muslim 1'. Data were collected via the Twitter streaming

Application Programming Interface (API), based on a manual inspection of the highest trending keyword following the event. The result was N=427,330 tweets in this case. The second incident was the *PunishAMuslimDay* event that took place on April 3rd 2018. The dataset was collected in the aftermath of a letter inciting others to commit violent and aggressive acts towards Muslims. This was given the name 'Anti-Muslim 2'. The collection spanned two weeks and resulted in N=919,854 tweets.

For the Anti-Semitic dataset we used a dataset collected by Ozalp *et al.* [249] who collected their dataset using the COSMOS platform[2]. They use a group of keywords which were agreed with CST [3] and are used for data collection: e.g. jew, jewish, antisemitic, nazis. These keywords are an collection of generic terms (Jew, Jewish, anti-Semitic, etc.). The data used for this analysis included tweets posted between 16/10/2015 and 21/10/2016 and were gathered in real time (this ensures that all tweets are collected). The raw dataset for the complete study period contained 31,282,472 tweets.

Chapter Five also establishes a link between Chapters Four and Five using the model developed in Chapter Four and applying it to detect cyberhate in the dataset collected in Chapter Five; the results are compared with human-annotated outcomes. Details of how the hateful networks were built and other insights are mentioned in Chapters Five and Six in the method section.

3.6 Summary

The framework is a pipeline that shows a map of how the methods are explained in each chapter. The general framework is divided into three sub-frameworks: cyberhate

[2] a free software tool that allows researchers to connect directly to Twitter's streaming Application Programming Interface (API) to collect real-time social media posts by specifying keywords

[3] The Community Security Trust is a British charity which exists to provide safety, security, and advice to the Jewish community in the UK.

classification, cyberhate characterisation and cyberhate disruption. The common process, namely the data collection, involved in the three frameworks, is then explained. The proposed framework will be discussed in the following four chapters.

Chapter 4

Hateful Content Classification

4.1 Introduction

As people increasingly communicate through web−enabled applications, the need for high−accuracy automated cyberhate detection methods has become much greater. Chapter 2 provided an insight into the studies related to the classification of cyberhate, explaining how recent studies have begun to interpret the effective features for machine classification of abusive language by focusing on how language is used to convey hateful or antagonistic sentiment. Several studies have shown how individuals with biased or negative views towards a range of minority groups are taking to the web to spread such hateful messages [195, 263]. Instances of cyberhate and expressions of racist views on social media have also been shown to be triggered by antecedent events, such as terrorist acts [65, 331]. A hate crime or bias-motivated crime occurs when the perpetrator of the crime intentionally selects the victim because of his or her membership of a certain group [300]. Hate speech is hereby understood to be inflammatory language that explicitly or implicitly targets an individual or group depending on protected characteristics.

Expressing discriminative opinions involves different language uses. For example, words might be used to convey intense dislike, such as *'hate them'*; moreover, to encourage violence, an inflammatory verb could be used, such as 'kill'.

There have been a number of attempts to automatically identify and quantify cyberhate by using different approaches, such as lexicons [131], syntactic [136] and semantic [183, 36] features, yet the limitation lies in classifying text that does not contain clearly hateful words which would have an impact on classification accuracy, *(e.g. send them home)*. While previous studies highlight the utility of methods capable of measuring semantic distances between words, such as embedding learning using individual words [106], and n-grams [243], this example requires an additional layer of qualitative context that sits above combinations of individual words.

Recent studies have begun to interpret the effective features for machine classification of abusive language by focusing on how language is used to convey hateful or antagonistic sentiment. 'Othering' - the use of language to express divisive opinions between the in-group ('us') and the out-group ('them') - has been identified as an effective feature [63]. The concept of 'othering' offers a potential candidate framework for the aforementioned qualitative layer capable of capturing the more subtle expressions of cyberhate such as the 'send them home' example. Anti-Hispanic speech might make a reference to border crossing or crime, anti-African American speech often references unemployment or single parent upbringing, and anti-Semitic language often refers to money, banking and the media. The use of stereotypes also means that certain types of language may be regarded as hateful even if no single word in the passage is hateful by itself [123].

This chapter illustrates a development created through a study carried out during this research: a novel method for cyberhate classification based around the use of 'othering language'.

The hypothesis is that *'linguistic features associated with othering language, as explained in Chapter 2, section 2.4.5, will provide an additional set of qualitative features that will improve the classification performance'*. Chapter 2, section 2.4.5 demonstrated that Intergroup Threat Theory (ITT) and Otherness are examples of psychological models that incorporate similarity and intergroup conflict and, therefore, could be useful for increasing our understanding of cyberhate.

Specifically, this chapter investigates whether the use of pronouns that refer to an ingroup (e.g. we, us) *co-occurring* with pronouns that refer to an outgroup (e.g. them, they) in the same post, will be indicative of divisive or antagonistic attitudes and therefore will improve machine classification of cyberhate. In this study, the co-occurrence of ingroup/outgroup pronouns is referred to as a *two-sided pronoun*. These were used to build a feature set that is referred to as an 'othering feature set', which is utilised to enrich the representation of text examples of cyberhate. These features are subsequently employed in combination with a paragraph embedding algorithm that has been shown in chapter 2 section 2.4.3 to outperform the BoW representation for cyberhate classification models and which has proved to be powerful when combined with other NLP features. Also, Schmidt *et al.* [281] identified that hate speech detection required sentence-level, rather than word-level, classification. Hence, this research chose the use of paragraph embeddings to classify hate speech.

The paragraph embedding algorithm infers semantic similarity between features to create a model that represents 'othering' language which is used for the purposes of cyberhate classification. Paragraph embedding algorithms aim to learn the semantic similarity of the proposed contextual features jointly with the rest of the text in the corpus. Samples that contain two-sided pronouns or pronoun patterns (e.g. verb-pronoun combinations) in a hateful or antagonistic context are aligned in similar feature 'spaces'. This increases the probability of the machine classification method labelling any new samples exhibiting these features as cyberhate. For example, the following sentence: (***We*** *want to hang* **them** *all*) contains the verb *hang"* and pronoun *them*, as well as the two-sided pronouns *we, them*. If the hypothesis is correct, and such features do indeed improve the context of the automated learning method, the sentence: (***We*** *need to get* ***them*** *out*) would be expected to be classified as cyberhate. This is not a sentence that would immediately flag as hateful by using existing classification methods but is an example of a sentence that would be identified by human annotators as a threat to individual groups and communities, and therefore needs to be considered when 'taking the social mood' following trigger events.

To benchmark the approach taken in this work, the results of different models that use state of the art classification algorithms and feature sets from the existing literature are presented and compared to the proposed method.

To the best of the author's knowledge, no study has yet used ITT theories in combination with the vector space feature to develop a feature set for improving the performance of machine learning for cyberhate.

The remainder of this chapter is structured as follows; section 4.2 presents the methods employed and explains the experimental steps. Subsequently, in Section 4.3, the classification results are presented and discussed. Finally, in Section 4.4, the contributions of this research are summarised.

4.2 Methods

This section reflects the design and development phase presented in Chapter 3. According to Chapter 3, the design and development phase aims to propose the 'othering feature set' to develop a classification framework. In addition, this section reflects the demonstration phase that has been mentioned in Chapter 3. The demonstration step aims to explain the methodology utilised for classifying hateful tweets by integrating the 'othering' concepts as a feature set, the aim being to develop a classification framework that takes advantage of the existence of the othering language in a hateful tweet. To achieve this novel othering layer within the machine classification framework, it was necessary to develop an othering feature set containing three components: (i) a constrained subset of dependency relationship labels extracted using probabilistic parse trees that were hypothesised to be representative of othering, (ii) more general parts of speech associated with othering including verbs (VB), nouns (NN) and adjectives (JJ) (e.g. send them home), and (iii) a list of English pronouns. Together, these capture a significant amount linguistic context to provide a focused set of othering features. This section details the process used to extract these features and how they were used in the automated machine classification approach. In particular, Section 4.2.1 introduces the datasets that were used in the study described in this chapter, and a statistical analysis of the use of othering language in these datasets. Sections 4.2.2 and 4.2.3 explain the extraction and the building of the othering feature set, while Section 4.2.4 explains the embedding feature extraction. Finally, Section4.2.5 shows the classification process.

4.2.1 Data

Datasets

To remind the reader, Table 4.1 shows the statistics of datasets used for the experiments in this chapter.

Table 4.1: Samples Numbers of Training And Testing Datasets

Training Dataset	**Number of Samples**
Davidson et al.[96]	5323 (hateful) 3161 (non-hateful)
Testing Datasets	Number of Samples
Religion	1901 (222 hateful)
Disability	1914 (53 hateful)
Race	1876 (73 hateful)
Sexual orientation	1803 (183 hateful)

Summary Statistics for Othering Language in the Datasets

To provide an initial justification for the research hypothesis on the use of two-sided pronouns to improve cyberhate classification, a corpus analysis of the datasets was conducted; it involved calculating the percentage of tweets from the Testing Dataset that included at least two pronouns. These features were found to be present in only 0.9% of non-malicious instances within the data and 17.6% of cyberhate instances.

Figure 4.1 shows the comparative occurrence of two-sided othering terms in both hateful samples and non-hateful samples among different types of cyberhate (test datasets).

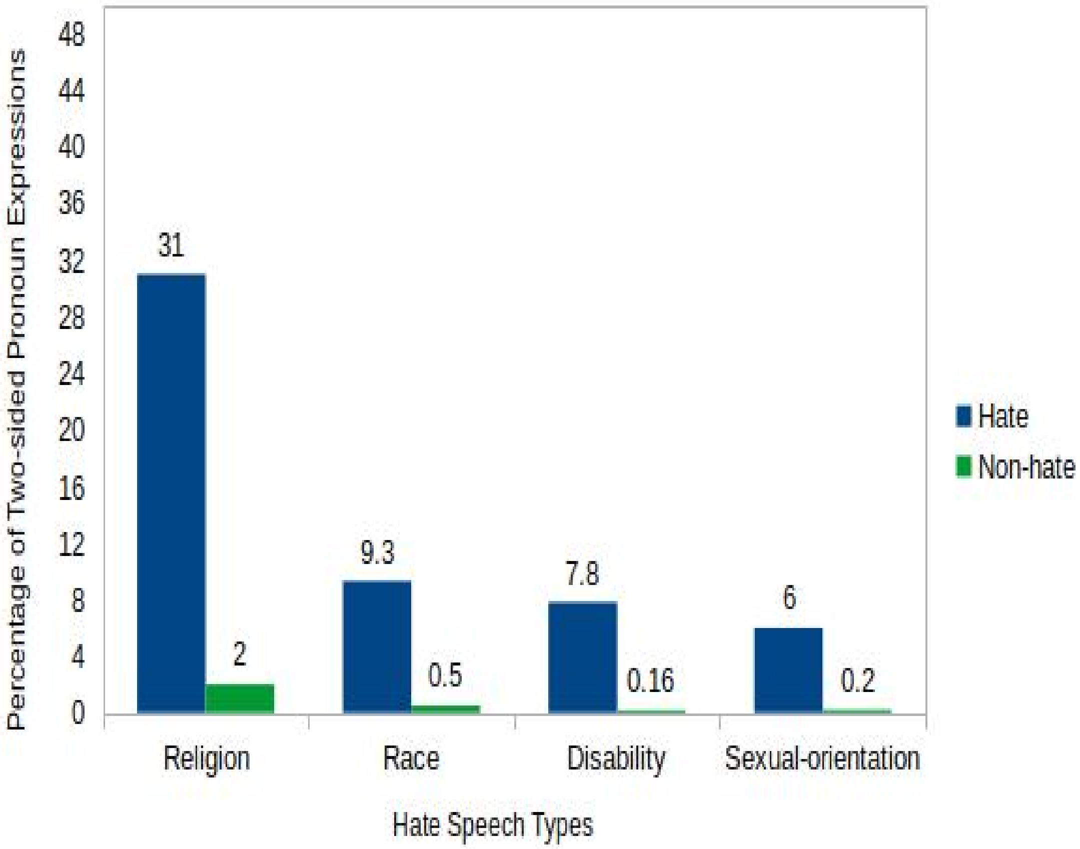


Figure 4.1: The use of two-sided othering for each hate speech type in both hateful and non-hateful samples in the testing dataset.

It can be seen that the Religion dataset contains the most frequent usage of the two-sided othering language.This follows from the findings of [65] who identified 'othering' language as a useful feature for classifying cyberhate based on religious beliefs. Also of note is that the percentage of two-sided othering was higher in the hateful annotated samples than non-hateful samples for all other types of hate speech, and to around the same level. In addition to that, Figure 4.2 illustrates the same comparison based on the results obtained using the Training Dataset which confirms again the much higher use of two-sided pronouns in the hateful samples.

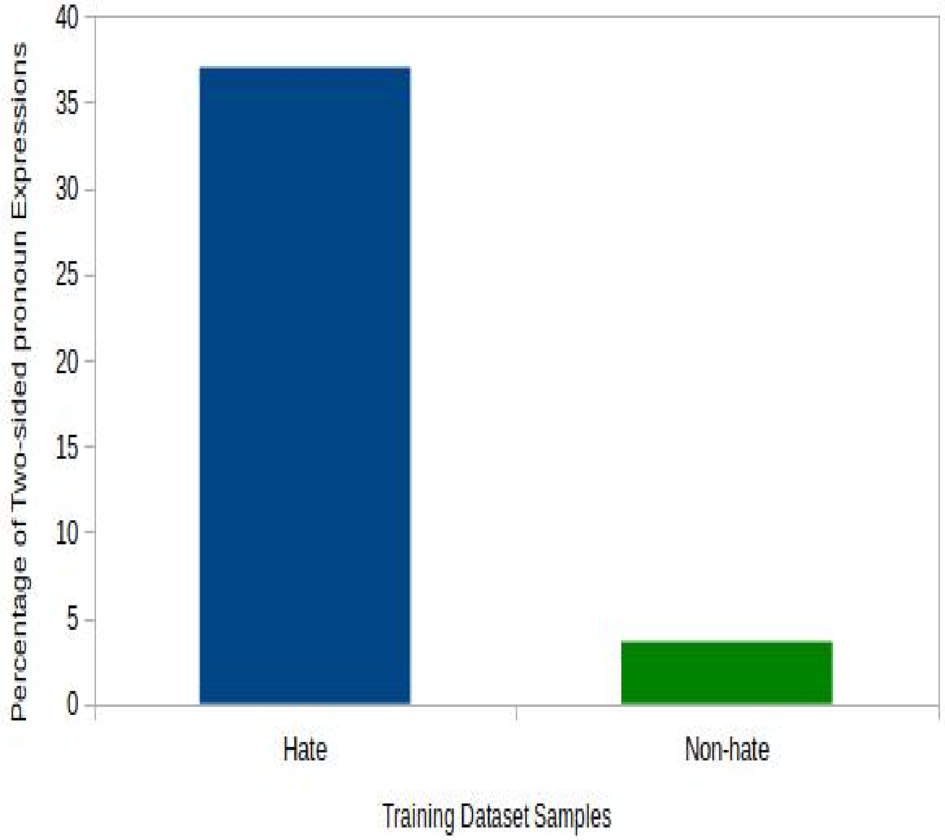


Figure 4.2: The use of two-sided othering for each hate speech type in both hateful and non-hateful samples in the training dataset.

4.2.2 Extracting Othering Terms

The first phase involved using the Training Dataset and analysing only the hateful samples. It is hypothesised that the use of pronouns that refer to an ingroup (e.g. we, us) *co-occurring* with pronouns that refer to an outgroup (e.g. them, they) in the same post, will be indicative of divisive or antagonistic attitudes and therefore will improve machine classification of cyberhate. In this chapter, the co-occurrence of ingroup/outgroup pronouns is referred to as a *two-sided pronoun*. Figure 4.3 presents an overview of linguistic features that can be used between different groups to distinguish themselves (the in group) from others (the out group).

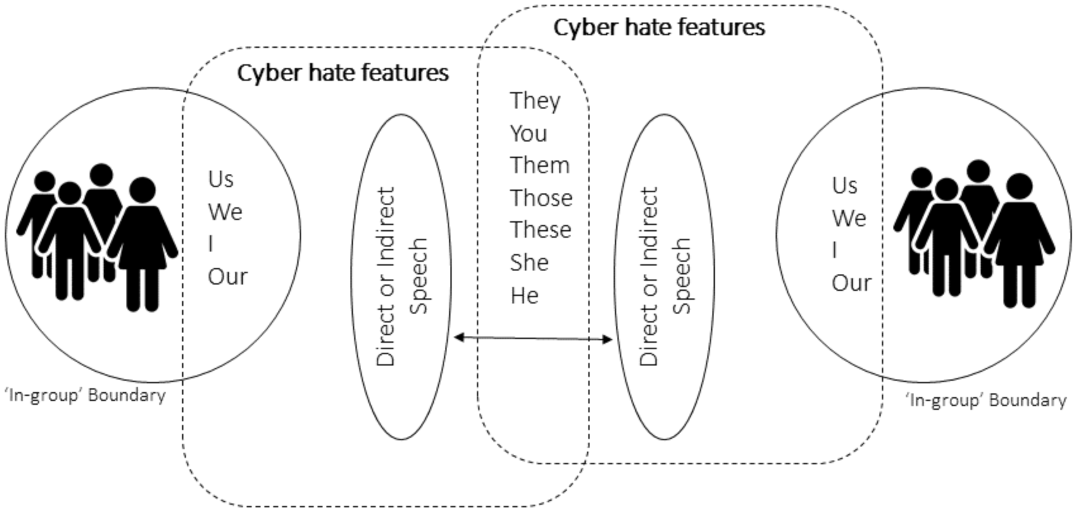


Figure 4.3: The boundary defined between two groups using the othering terms, and the space between the boundary shows how the negative text could be defined.

The figure illustrates how pronoun terms from one side (us, we, our, etc.) draw the boundary between the in group by referring to the out group (we, they etc). Using 'us' and 'them' in the same context; these very pronouns highlight the distinction between the groups, as the first-person plural pronoun, 'us' places the speaker within a group with e.g. a shared identity. By contrast, 'them' is the third person plural, which is used to refer to people at a distance. All samples where two-sided othering was present - i.e. where at least two pronouns were used, were identified. An example tweet of two-sided othering is *"send **them** home **we** are fed up"*. Here, the tweet contains two pronouns that draw boundaries between two different groups. Any samples without at least two pronouns were discarded, as were the repeated samples. Using this subsample, a Typed Dependency parser was utilised, which was explained in Chapter 2, to transform the text to provide co-occurring words with a probabilistically derived linguistic label, specifically, the Stanford Typed Dependency Parser. The Stanford Typed Dependency Parser provides a representation which was designed to provide a simple description of the grammatical relationships in a sentence that can easily be understood and effectively used by people without linguistic expertise who want to extract textual relations. In particular, rather than the phrase structure representations that have long dominated in the computational linguistic community, it represents all the sentence relationships uniformly as typed dependency relations [100]. Parsing a text into a Typed Dependency relation relies on the use of a Dependency Parser. This package is a Java implementation of a probabilistic natural language parser that requires Java 8+ to be installed. The parser also requires a reasonable amount of memory[1]. in order for the dependencies to use the Linux- Ubuntu command line to train models and to parse text. See the following Linux- Ubuntu command:

```
java -mx1g -cp ""*""
 ↪   edu.stanford.nlp.parser.lexparser.LexicalizedParser
 ↪   -outputFormat ""typedDependencies"" -sentences
 ↪   newline -encoding utf-8 -model
 ↪   stanford/nlp/models/lexparser/englishPCFG.ser.gz
 ↪   InputFile.txt > OutputFile.txt.out
```

The options used for the command line above are explained in Table 4.2:

Table 4.2: Description of the options that have been used for setting the Typed Dependency model.

Option	Description
-outputFormat	To obtain dependency formatting options for the parse tree, we add *typedDependencies* in the -outputFormat option
-sentence	Pars the text sentence by sentence Pars the text sentence by sentence
-model	Path to a model file. During training our model, we used a lexical parser: *englishPCFG.ser.gz* model.
-outputFile	Specify the output file

The resulting text for parsing the example ***"send them all home we do’nt want them in our country"*** as follow:

[1] at least 100MB, to run a PCFG parser on sentences up to 40 words in length; typically, around 500MB of memory is needed to parse similarly long typical-of-news wire sentences using the factored model

root(ROOT-0, send-1)

iobj(send-1, them-2)

det(home-4, all-3)

obj(send-1, home-4)

nsubj(do-6, we-5)

acl:relcl(home-4, do-6)

advmod(send-1, nt-8)

dep(send-1, want-9)

obj(want-9, them-10)

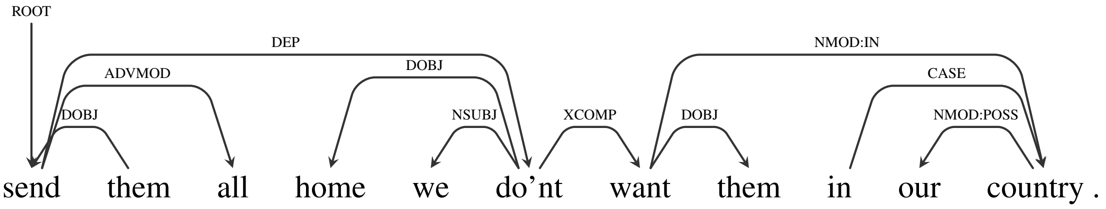
case(country-13, in-11)

nmod:poss(country-13, our-12)

obl:in(want-9, country-13)

Figure 4.4 shows the linguistic labels associated with each word in a sample sentence. Word order within a sentence is preserved according to type-dependency and provides a feature for classification as well as the syntactic relationship between words.

Figure 4.4: The dependencies that display the linguistic labels associated to each word in an example phrase.



The Stanford Parser returns 51 different linguistic labels [2]. The dependencies are all binary relations: a grammatical relation holds between a governor (also known as a regent or a head) and a dependent. In the previous example, the parser produced seven dependency relationships, distributed over ten instances. To provide a specific focus on

[2] https://nlp.stanford.edu/software/dependencies_manual.pdf

othering language, only six types of dependency relationships were retained: *nsubj*, *dobj*, *nmod*, *det*, *advmod* and *compound*. The remaining dependency modifiers were discarded. The rationale for preserving these modifiers is as follows. The *nsubj* label captures the syntactic subject or proto agent in a sentence (i.e. the active agent). Examples include 'Muslims caused' and 'they inflicted'. *dobj* concerns the direct object of a verb phrase and has a high probabilistic likelihood of capturing relationships between verbs and nouns, pronouns and determiners in the same phrase (e.g. send and them). *nmod* is likely to identify nominal modifiers for nouns, for instance 'all gays' or 'womens place'. The *det* captures the relationship between nominals and their determiner (e.g. 'these terrorists'). *advmod* captures adverb modifiers (e.g. where we see 'home' we may also see 'send'). The *compound* will identify compound verb phrases including verb and adjective compounds such as 'send back' or 'kill black'.

As a worked example of how this method is expected to capture othering, the translation of the text in Figure 4.4 becomes *nsubj(want-7, we-5), dobj(send-1, them-2) det(home-3, all-4), nmod:poss(country-11, our-10)* and the remaining relationships would be discarded. We are now capturing distinctive othering features that co-occur in the same sentence. Despite none of these words being clearly antagonistic or offensive on their own, together they provide a greater contextual feature for machine classification to detect these unseen samples using similar phrasing.

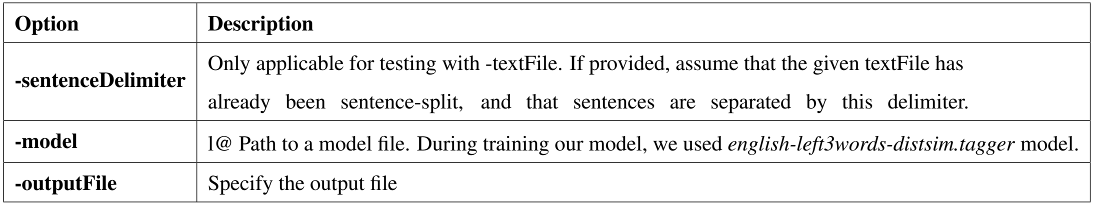
4.2.3 Building the Othering Feature Set

To complement the dependency relationship features, part-of-speech (POS) tagging to the hateful samples that included at least two pronouns was also applied. Part-Of-Speech was defined in Chapter 2 as the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context. Once again, the set of labels was refined to include those most likely to represent othering and retained only words tagged as nouns (NN), adjectives (JJ), verbs (VB) and adverbs (RB). POS also requires Java 8+ to be installed and uses the Java package for implementation. POS was obtained using the following Linux-Ubuntu command line:

```
java -cp ""*""
 ↪  edu.stanford.nlp.tagger.maxent.MaxentTagger
 ↪  -sentenceDelimiter newline -model
 ↪  stanford/nlp/models/pos-tagger/
english-left3words/english-left3words-distsim.tagger
 ↪  InputFile.txt -outputFile  OutputFile.txt
```

The options used for the command line above are explained in Table 4.3:

Table 4.3: Description of the options that have been used for setting the POS model.



Option	Description
-sentenceDelimiter	Only applicable for testing with -textFile. If provided, assume that the given textFile has already been sentence-split, and that sentences are separated by this delimiter.
-model	1@ Path to a model file. During training our model, we used *english-left3words-distsim.tagger* model.
-outputFile	Specify the output file

The resulting text for the example *send them all home we do’nt want them in our country* as follow:

send/VB them/PRP all/DT home/NN we/PRP do/VBP’/”nt/RB want/VB them/PRPin/IN our/PRP country/NN

The POS labels themselves were removed to leave only words. These POS words, the dependency relationship features and a list of all English pronouns were then concatenated into a triple that formed the basis of an othering feature set - a novel concatenation of a range of grammatical and linguistic features extracted from a human annotated data set of hateful and antagonistic texts. To reduce noise, all the tweets that contained a single word were removed. This process resulted in a dataset of 975 rows. As the othering feature set was built using annotated hateful samples, it was expected that all the entries could contribute to the learning process. Figure 4.5 shows the process of extracting the 'othering feature' from each tweet containing two-sided pronouns. Algorithm 4.1 illustrates the steps of building the othering vectors feature set. Figure 4.6 shows the process of training the model and Figure 4.7 illustrates the testing phase.

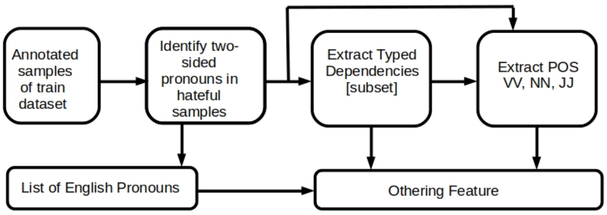

Annotated samples of train dataset
Identify two-sided pronouns in hateful samples
Extract Typed Dependencies [subset]
Extract POS VV, NN, JJ
List of English Pronouns
Othering Feature

Figure 4.5: Othering Feature Extraction

Three types of training datasets were examined: the first training dataset contains all the non-hateful and hateful instances that contain two-sided pronouns in an unprocessed form - i.e. raw text - named **Unprocessed Training Dataset**. The second training dataset contains all the non-hateful instances transformed into Typed Dependency representation, plus the othering feature set to represent hateful instances. This is named **Proposed Feature Set 1**. The third training dataset merged the first training dataset and the second dataset which is named **Proposed Feature Set 2**. Examples of each input are as follows: (1) *'Send them all home we don't want them in our country'*; (2) *'Row0: [(them,we,our) + nsubj(want-7, we-5), dobj(send-1, them-2) det(home-3, all-4), nmod:poss(country-11, our-10) + send,want,home]'*. These features then become the input for the Paragraph2Vec algorithm.

Algorithm 4.1 Othering feature set

INPUT: Annotated Training Dataset

OUTPUT: Othering Feature set

```
procedure Generation of the othering feature set
for  each samples in Training Dataset do
    Identify tweets containing two sided pronoun
    for each sample containing two sided pronoun do
        Extract all the pronoun ($P$)
        Extract Typed Dependency ($TD$)
        Extract POS ($POS$)
        Append ($P$,$TD$,$POS$)
    end for
end for
Generate the othering feature set
End procedure
```

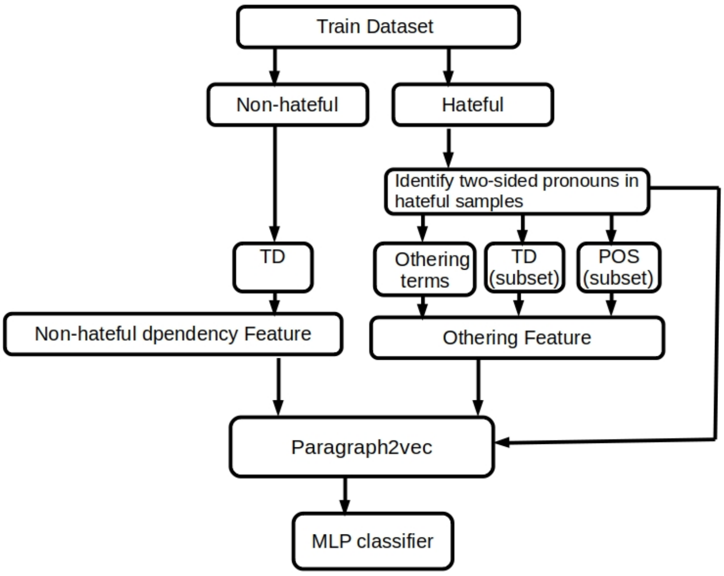


Figure 4.6: Model Training Workflow

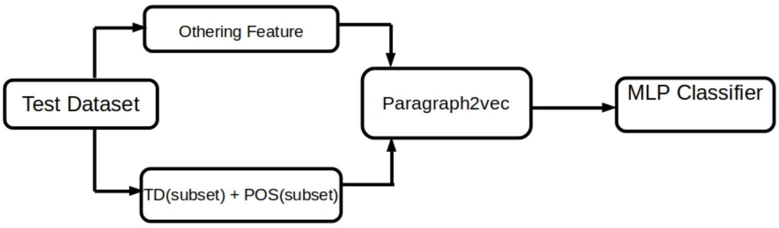


Figure 4.7: Model Testing Workflow

4.2.4 Embedding Feature Extraction

At this stage, it was necessary to identify a suitable method for utilising the features extracted through the othering feature set; these could have been used as raw features for classification, but to provide further refinement, embedding learning was employed to capture the relative ‘distance’ between these features in a cyberhate context. Learning vector representations allow each feature to be plotted in such a way that the numeric distances between features based on their use in a context can be calculated. A common example of this is that the distances between ‘man’ and ‘king’ would be similar to that of ‘woman’ and ‘queen’. Therefore, relationships between words (i.e. ‘man’ and ‘king’), and context (i.e. ‘king’ is to ‘queen’ as ‘man’ is to ‘woman’) can be identified. With two-sided pronouns, it is assumed that the method would benefit from this approach to extract the semantic ‘meaning’ of othering features and learn these jointly across the hateful and non-hateful texts to provide context for term use - ultimately with the aim of these features being better able to support the machine classification of both.

Various methods have been proposed to learn vector embedding representations. As mentioned in Chapter 2, Word2Vec and Paragraph2vec have been proposed for building word/paragraph representations in low-dimensional vector space [219]. In the Word2Vec model, *words* are represented in continuous space where semantically similar words have a high similarity measure in that space. In the Paragraph2vec model, each sentence is mapped to a unique vector, and every word included in the sentence is also mapped to a unique vector.

In the Paragraph Vector-Distributed Memory component (PV-DM), the sentence acts as a memory that remembers the missed word in the current context of the sentence.

The paragraph vector and word vectors are averaged or concatenated to predict the next word in a context [192]. The previous method considers the concatenation of the paragraph vector with the word vectors to predict the next word in a text window. In (PV-DBOW), The paragraph vector can be further simplified when we use no local context in the prediction task. Model Learning the embedding using Paragraph Vector-Distributed Bag-of-words (PV-DBOW) ignores the context words in the input but forces the model to predict words randomly sampled from the paragraph in the output.

In the context of this study, this means each tweet is fed into the embedding learning methods as if it were a sentence, and each feature of the tweet derived in the othering feature extraction phase becomes a part of the sentence embedding.

Both datasets 1 and 2 were transformed into paragraph embeddings using Paragraph2Vec for training and testing purposes. Note, Word2Vec was also implemented for sentence classification, which resulted in poor classification results; therefore, the decision was made to discard the use of Word2Vec and use only Paragraph2Vec.

Distributed Memory (PV-DM) vectors were learned using the Gensim[3] implementation of distributed representations of the sentence (tweet) [192]. According to Mikolov *et al.* [220], the distributed memory model is consistently better than the Distributed Bag of Words (PV-DBOW). To find the best implementation for the data, experiments with both were carried out. The results showed that distributed memory performed better in learning feature vectors for the dataset. The Gensim package using the Python language runs on a Linux- Ubuntu Operating System. An example code of model build and training as follow:

```
model = Doc2Vec(min_count=0, window=2, size=600, workers
    =8,dm=0)
model.build_vocab(sentences.to_array())
model.train(sentences.sentences_perm(),total_examples=
    model.corpus_count, epochs=model.iter)
```

[3] https://radimrehurek.com/gensim/models/Doc2Vec.html

Table 4.4: Definitions of the parameters that have been used for setting the Paragrap2vec model.

Parameter	**Definition**
dm	Defines the training algorithm. If dm=1 then'distributed memory' (PV-DM) is used. Otherwise, distributed bag of words (PV-DBOW) is employed.
Windows_Size	The maximum distance between the current and predicted word within a sentence.
Vector_Size	Dimensionality of the feature vectors.

The parameters used here are summarised in Table 4.4.

Small window sizes were used because, according to Levy *et al.* [200], a window of size 5 is commonly used to capture broad topical contents, whereas smaller windows (e.g. *k=2* windows) contain more focused information regarding the target word.

For example, for k = 2, the context around the target word w comprises w- 2, w- 1, $w+1$, $w+2$. These become the features used for learning distances between the target word and its surrounding context. The larger the window, the broader the context. The final output is a vectorised dataset that is used as a feature set for feeding into a machine classification approach. As the othering layer was expected to assist in improving the performance of machine classification of cyberhate, a more focused approach seemed logical, given it would be these nuanced othering terms in a smaller window that would likely lead to improvements in classification. Various window sizes including 100, 300, 600, 800 and 1000 dimensions and k = 2, 3, 5, 6 and 10 were experimented with. The performance of each was recorded and the best performing configuration was reported on - which was 600 dimensions and *windows = 2*. Fixed embedding was used because the datasets are not associated with time, which would require grouping the data into time bins and training the embeddings separately on these bins [181]. Once the vector representations had been learned, the vector was joined with its original human-assigned label, assigning the label 0 to non-hateful samples and 1 to hateful samples, and then, these were used to test the machine classifier.

4.2.5 Machine Classification

Machine learning models take samples of labelled text to produce a classifier that is able to detect hate speech based on labels annotated by content reviewers. To identify or classify user-generated content, text features indicating hate must be extracted. The usefulness of a novel classifier will be determined by comparing it with state-of-the-art classifiers.

Comparing the 'Othering' Classifier to Baselines

Several classification approaches were examined, drawing on state-of-the-art related cyberhate research to determine the overall improvement when using the novel othering feature set. The candidate methods included:

- **(Baseline 1)** - Support Vector Machines (SVM) and Random Forests (RF) combined with Bag of Words (BoW), n-gram, and Typed Dependency features, as used in [62, 63]. The SVM parameters were set to normalise data, use a gamma of 0.1 and C of 1.0 and we employed radial basis function (RBF) kernel and the Random Forest (RF) iteratively to select a random sub-sample of features in the training stage and train multiple decision trees before predicting the outputs and averaging the results which maximise the reduction in classification error [58]. The Random Forest algorithm was trained with 100 trees.
- **(Baseline 2)** used a Logistic Regression (LR) classifier with Paragraph2Vec feature extraction for joint modelling of comments and words, as used in [106]. They used the CBOW model as a component of paragraph2vec [220], which tries to predict the central word based on the surrounding words, as well as the user comment the words belong to.
- **(Baseline 3)** used Vowpal Wabbit's regression model and different NLP features with Paragraph2Vec and Word2Vec used for feature extraction, as applied by

Nobata *et al.* [243]. Their features can be divided into four classes: N-grams, Linguistic (e.g. length of comment in tokens, average length of word and number of punctuation items), Syntactic (e.g. POS) and Distributional Semantics (e.g.Paragraph2Vec and Word2Vec).

- **(Baseline 4)** included a CNN model in combination with Word2Vec embedding. Training was performed in batches of size 128 for CNN as introduced in [125] and the 'adam' optimiser for CNN was used. The model was configured with three max-pooling layers, as introduced in [125]. A max-pooling layer captures the most important latent semantic factors from the tweets. The output layer used softmax to calculate the class probability distributions for each tweet and assigned each tweet the class that obtains the maximum probability value.

- **(Baseline 5)** used Gradient Boosted Decision Trees (GBDTs) in combination with Long Short-Term Memory (LSTMs) feature extraction (not as a classifier), and random embeddings, as published by Badjatiya *et al.* [36]. They used the LSTM model to capture sequence-based features. The LSTM model has a single layer of LSTM units and all of the words in the corpora were initialised with random values. The output dimension size of the LSTM layer was 100. A sigmoid layer was built on the top of the LSTM layer to generate predictions. The input dropout rate and recurrent state dropout rate were both set to 0.2. In each iteration of the bootstrapping process, the training of the features and classifier runs for 15 epochs.

- **(Baseline 6)** was a modified CNN classifier with a Gated Recurrent Units (GRU) layer which was applied to learned Word2Vec embeddings as introduced by Zhang *et al.* [354]. They integrated a GRU layer with a CNN classifier to capture long range dependencies in tweets, which may play a role in hate speech detection. A GRU layer takes input from the max pooling layer. This treats the features as time steps and outputs 100 hidden units per time step. Compared to LSTM, the key difference in a GRU is that it has two gates (reset and update gates) whereas an LSTM has three gates (namely input, output and forget gates). Thus, GRU is a simpler structure with fewer parameters to train. In theory, this makes it faster to train and better for generalising on small data, while empirically it is shown to achieve comparable results to LSTM [83].

- (**Baseline 7**) used an LSTM classifier with random embedding which was introduced in [36] but did not produce an improvement on the use of GBDTs in their results. However, in this study, the aim was to reveal the effectiveness of using an LSTM model as a standalone classifier on the proposed feature set, and the parameters were set in the same way as for [36]. For CNN, LSTM and CNN+GRU models, the feature extraction phase all resulted in paragraph level vectors being extracted for each sentence (tweet). This means that all of the baselines (except baseline 1) have a vectors space feature for classifier inputs, which makes them comparable to Paragraph2Vec as used in the othering feature set.

In addition to the state-of-the-art classification methods, the use of our othering feature set (pre-processed with Paragraph2Vec) along with a Multilayer Perceptron (MLP) classifier [341] is proposed. Multilayer feed-forward networks can provide competitive results for sentiment classification and factoid question answering [159]. As mentioned in Chapter2, MLP is a feed-forward artificial neural network model which maps input datasets on an appropriate set of outputs. MLP consists of multiple layers of nodes in a directed graph, with each layer being fully connected to the next layer [129]. This thesis chooses to examine the MLP classifier as it a classical form of the neural network (NN) classifiers which showed an improvement in detecting cyberhate compared to non-neural network classifiers. Fortuna *et al.* [119] reported that the MLP classifier achieved better performance in detecting higher hateful samples than SVM, LR and RF. Also, the MLP classifier has not been used combined with embedding features in previous studies. To set an MLP classifier, various parameters should be set to configure the classifier network, Table 4.5 defines these parameters.

MLP parameters were set experimentally by setting the initial number of the hidden

Table 4.5: MLP parameters

Parameter	Definition
Node (units)	node, also called a neuron or Perceptron, is a computational unit that has one or more weighted input connections, a transfer function that combines the inputs in some way, and an output connection.
Layer	Nodes are then organised into layers to comprise a network
Iteration	Maximum number of iterations this determines the number of epochs (how many times each data point will be used).

layers to 1 and increasing the number of the hidden layers to improve performance through trial and error [148]. In this case, two hidden layers with five hidden connected units achieved the best performance for the vectors with 200 iterations. The implementation of the MLP classifier was achieved using the SKlearn package in the Python language:

```
from sklearn.neural_network import MLPClassifier
classifier = MLPClassifier()
```

All these approaches were implemented on the unprocessed training dataset, on the proposed feature set 1, and on the proposed feature Set 2. Then, to determine the effectiveness of each individual model in classifying cyberhate, they were cross-validated on an individual basis across all the input feature sets. For each cross-validation fold, the paragraph embeddings were re-trained. Throughout the results section, correct classifications of non-hateful samples are referred to as 'true negatives', and correct classifications of hateful samples as 'true positives'. Therefore, a misclassification of non-hate classified as hate is a false positive (FP), and a misclassification of hate classified as non-hate is a false negative (FN).

Testing the Othering Classifier

This experiment refers to testing the model's ability to correctly classify new, unseen data, drawn from the same distribution as that used to create the model. At this step, the best performing classifiers would be tested on four subsets within the Testing Dataset as described earlier: (i) Religion, (ii) Disability, (iii) Race and (iv) Sexual orientation. The aim of this step was to examine the generality of the 'Othering' concept across different types of cyberhate. The generalised cyberhate classifier should maintain good performance when applied to new data. The following section shows the results of the above experiment and the related discussion.

4.3 Results and Discussion

The evaluation phase of the Data Science Research Methodology (DSRM) outlined in Chapter 3 is reflected in this section. During this phase, the evaluation experiments aim to measure if the linguistic features associated with othering language provide an additional set of qualitative features that will improve performance in terms of classification.

4.3.1 Quantitative Results

The first set of experiments included applying a wide range of models from previous studies, which were summarised as baselines in the previous section. This allows for a comparison of the best performing models for cyberhate classification with the proposed 'othering' feature set. The second set of experiments involved testing the best performing models from the first phase on completely unseen data to test model generality over four types of hate speech.

Testing against the state of the art

Table 4.6: Machine classification performance for the cyberhate classifiers based on training dataset .

			Unprocessed Training Dataset **(The raw dataset)**				**Proposed Feature set 1** **(Othering Feature set)**				**Proposed Feature set 2** **(The raw dataset + othering feature set)**			
Classifiers	**cl**	**sample**	**P**	**R**	**F**	**AUC**	**P**	**R**	**F**	**AUC**	**P**	**R**	**F**	**AUC**
Baseline model 1: n-Gram words (1-5) with 2,000 features +n-Gram typed dependencies +hateful term [63]	SVM+ RF	Hateful	0.48 FN=407	0.58 FP=610	0.52	0.67	0.30 FN=311	0.68 FP=1560	0.41	0.56	0.28 FN=209	0.78 FP=1891	0.42	0.57
		Neutral	0.80	0.86	0.83		0.50	0.83	0.63		0.40	0.85	0.54	
Baseline model 2: Comment Embedding [106]	LR	Hateful	0.93 FN=118	0.88 FP=61	0.91	0.94	0.87 FN=253	0.74 FP=110	0.78	0.89	0.96 FN=10	0.98 FP=40	0.97	0.97
		Neutral	0.98	0.96	0.97		0.95	0.92	0.94		0.98	0.99	0.99	
Baseline model 3: N-grams+ linguistic+ syntactic+ word and comment Embedding [243]	VR	Hateful	0.93 FN=54	0.94 FP=66	0.92	0.95	0.98 FN=24	0.97 FP=9	0.98	0.99	**0.99** **FN=4**	**0.98** **FP=9**	**0.99**	**0.99**
		Neutral	0.97	0.98	0.98		0.99	0.99	0.99		1.00	1.00	1.00	
Baseline Model 4: Word2vec [125])	CNN	Hateful	0.88 FN=27	0.97 FP=131	0.92	0.93	0.83 FN=15	0.98 FP=196	0.90	0.91	0.90 FN=11	0.98 FP=101	0.94	0.95
		Neutral	0.95	0.99	0.97		0.93	0.99	0.90		0.96	0.99	0.98	
Baseline Model 5: Word2vec +LSTM [36]	GBDT	Hateful	0.77 FN=258	0.73 FP=215	0.75	0.84	0.89 FN=17	0.98 FP=102	0.94	0.95	0.95 FN=34	0.96 FP=47	0.97	0.97
		Neutral	0.93	0.91	0.92		0.96	0.99	0.98		0.98	0.99	0.98	
Baseline Model 6: Word2vec [355]	CNN+ GRU	Hateful	0.87 FN=75	0.92 FP=133	0.90	0.92	0.92 FN=25	0.97 FP=30	0.95	0.98	0.89 FN=10	0.99 FP=115	0.94	0.94
		Neutral	0.95	0.97	0.96		0.99	0.99	0.99		0.96	0.99	0.97	
Baseline 7:Word2vec [83]	LSTM	Hateful	0.00 FN=975	0.00 FP=3161	0.00	0.00	0.36 FN=365	0.59 FP=1003	0.45	0.60	0.45 FN=158	0.83 FP=974	0.59	0.69
		Neutral	1.00	0.76	0.86		0.68	0.84	0.75		0.69	0.93	0.79	
Paragraph2vec	MLP	Hateful	0.95 FN=21	0.98 FP=51	0.96	0.97	**0.99** **FN=5**	**0.99** **FP=3**	**1.00**	**0.99**	**.99** **FN=7**	**0 0.99** **FP=5**	**0.99**	**0.99**
		Neutral	0.98	0.99	0.98		1.00	1.00	1.00		1.00	1.00	1.00	

The results are shown in Table 4.6, in which the first column represents a summary of the baseline models ([63], [106], [243], [36], [125], [354] and [83]). The last row in the table contains the results of the novel approach taken in this research. In this phase, all methods were evaluated using ten-fold cross validation. This method has previously been used for experimentally testing machine classifiers for short text [308] [63]. It functions by iteratively training the classifier on feature vectors from 90 percent of the annotated dataset and classifying the remaining 10 percent as 'unseen' data, based on the features evident in the cases it has encountered in the training data. It then determines the accuracy of the classification process and moves on to the next iteration, finally calculating the overall accuracy. In the training phase, the F-measure was used as the main comparison metric, given it controls for false positives and false negatives and a lack of balance in the dataset. Also, AUC was used for error analysis. AUC provides an aggregate measure of performance across all possible classification thresholds, ranging from 0 to 1; a model with predictions that are 100% incorrect has an AUC of 0.0, and one with predictions that are 100% correct has an AUC of 1.0.

The results presented in Tables 4.6 and 4.7 are for the cyberhate class only, as the main interest here is in the improvement of cyberhate classification.

As mentioned previously in section 4.2.3, the experiments were conducted on three input datasets: (i) **unprocessed training dataset** which was just the raw text and reflected the (actual) implementation of the applied models; (ii) **proposed feature set 1**; and (iii) **proposed feature set 2** which was a combination of the unprocessed training dataset and proposed feature set 1. This follows the method proposed in Sections 3.5.1 to 3.5.3.

From Table 4.6 we firstly notice that from baseline 1 to baseline 2 there was a large reduction of FPs and FNs among the three feature sets. We suggest this is likely due to the use of semantic learning for features extraction. This also confirms that there is no bias in the proposed datasets as noticed by [42]. From baseline 2 to baseline 3, the reduction of FNs (but not necessarily FPs) is clear. We suggest this is likely a result of using n-grams, linguistic (e.g. length of tokens) and syntactic (e.g. POS) features. The combination of baseline 3 with the third feature set produced the lowest number of FPs, detecting 99% of the cyberhate samples. This suggests that the extra features (n-gram and linguistics), which were applied by the baseline 3 model, improved the process of hate speech classification, compared with the baseline models, which already use the semantic features.

However, applying the neural network models proposed in baseline 4 [125], baseline 5 [36], baseline 6 [355] and baseline 7 [83] did not result in a significant increase in the detection of cyberhate within the three datasets compared to baseline 3. The CNN [125] and LSTM models [36] were unable to improve on this using the unprocessed feature set (the raw feature set) or the proposed feature sets. This is likely due to the length and sparsity of short texts. Furthermore, RNN models including LSTM features extraction, which resulted in the best performance in [36] and the GRU layer, which was added to the CNN model to achieve best performance by [356] showed weak performance compared to the other classifiers. Another possible explanation for this is their use of Word2Vec feature extraction which, empirically, has not produced better semantic learning than Paragraph2Vec for the datasets in this thesis (see Section 4.2.3). The last row in Table 4.6 shows the results of training the proposed feature sets using the Paragraph2Vec model for feature extraction and the MLP classifier. Despite the two proposed feature sets producing higher FN than baseline 3 on the proposed feature set 2 in this research by 1 and 3 extra FN samples, they show a reduction of FP by 6 and 4 non-hateful samples. The 'othering language' features here are working on a par with baseline 3 that uses a range of text pre-processing methods.

Overall, the experiments show that the 'othering language' features alone (proposed feature 1) are valuable for enhancing the detection of the hateful instances for the majority of (but not all) baselines' classifiers. For all the baselines' classifiers (except baseline 1), including the raw dataset with the 'othering language' feature set (proposed feature 2) led to the best overall performance (best f-measure) improvement by 2%-59%. However, for the approach taken in this study, using the 'othering language' features alone (proposed feature 1) produced the best performance - slightly better than (proposed feature 2).

This shows the ability of the 'othering' narrative to be an essential part of hate speech detection, suggesting that using the embedding representation of the 'othering feature set' would provide better context for the classifier beyond using BOW, n-grams, linguistic and syntactic features.

To evaluate the generality of the best performing models (the proposed model and baseline 3) - and to stress-test the range of linguistic features used in baseline 3, as well as the othering method - they were tested using unseen datasets. The three best performing classifiers (shown in bold font in Table 4.6) were the candidate models for further testing because they resulted in the highest F-measure and AUC >= 0.99 (also the lowest number of both FPs and FNs ranging between 0 and 10). The three best models were referred to as the 'Comprehensive-classifier', 'Othering-classifier', and 'Othering+raw-classifier', respectively. The term Comprehensive-classifier refers to applying the baseline 3 model to the proposed feature set 2, referred to as *'comprehensive'* because a wide range of features was applied. The Othering-classifier refers to the proposed application of Paragraph2Vec feature extraction and the MLP classifier to the proposed feature set 1 (othering feature set), and Othering+raw-classifier refers to the proposed application of Paragraph2Vec feature extraction and the MLP classifier to the proposed feature set 2 (othering feature set and raw data set).

Testing the Othering Classifier

The second set of experiments carried out in this research involved testing the proposed model on unseen datasets. The training phase (the previous experiment) produced evidence to suggest that including othering language features in predictive models for cyberhate speech (in short informal text, such as Twitter posts) would improve the classification performance. The second phase aimed to determine the possibility of developing a more generalised model for cyberhate detection. A key finding from previous research is that, compared with using hateful terms alone, the inclusion of features capable of detecting othering language in the classification of religious cyberhate reduced false negatives by 7%. Additionally, Nobata *et al.* [243] found that computing feature embeddings when combined with the standard NLP features showed the effectiveness of improving the performance of cyberhate classification. The utility of the learned vectors in the classification of cyberhate was validated by reporting precision (P), recall (R), F-measures (F) and, the area under the curve (AUC).

The results of the experiments show how the use of othering features alongside embeddings to train the classifier enables a new level of hate speech detection, in the form of othering-level feature embeddings. The best three classifiers from previous experiments were tested on four unseen cyberhate datasets.

- **Religious hate:** as shown in Table 4.7, for religion, the Othering+raw-classifier resulted in the highest recall (R)=0.99. This means it is the best at detecting hate.

Table 4.7: Machine classification performance for cyberhate classifiers based on unseen testing datasets .

	Test Data Sets															
	Religion				**Disability**				**Race**				**Sexual-orientation**			
Trained Models	**P**	**R**	**F**	**AUC**	**P**	**R**	**F**	**AUC**	**P**	**R**	**F**	**AUC**	**P**	**R**	**F**	**AUC**
Baseline Model 3 + **Proposed Feature set 2** **(Comprehensive-classifier)**	**0.71** **FN=12**	**0.95** **FP=84**	**0.81**	**0.94**	**0.74** **FN=16**	**0.68** **FP=12**	**0.71**	**0.86**	0.30 FN=8	0.89 FP=163	0.43	0.64	**0.31** **FN=3**	0.98 FP=398	0.47	0.65
Paragraph2vec+ **Proposed Feature set 1** **(Othering-classifier)**	0.44 FN=12	0.95 FP=260	0.60	0.71	0.94 FN=35	0.31 FP=1	0.47	0.96	0.79 FN=4	0.95 FP=18	0.86	0.89	0.46 FN=40	0.78 FP=168	0.58	0.71
Paragraph2vec+ **Proposed Feature set 2** **(Othering+raw-classifier)**	0.54 FN=2	0.99 FP=187	0.69	0.76	1.00 FN=23	0.55 FP=0	0.71	0.99	**0.84** **FN=3**	**0.95** **FP=13**	**0.89**	**0.92**	**0.61** **FN=24**	**0.86** **FP=99**	**0.72**	**0.80**

It resulted in the lowest FNs (lack of detection of the hateful samples). However, it also had poorer results than the comprehensive classifier in terms of precision (P)=0.54 and F-measure (F)=0.69.

The lowest FPs (non-hate classified as hate) was achieved by the Comprehensive-classifier, which resulted in Precision (P)=0.71, Recall (R)=0.95 and F-measure (F)=0.81. Therefore, this model is considered to be a more balanced classifier. We can see evidence for this in that the Comprehensive-classifier achieved a higher AUC (0.94) compared to Othering+raw-classifier. For religious hate, having the additional features in the comprehensive classifier (n-grams, linguistics, etc) helps balance the classifier - but not having them (i.e. in the othering+raw classifier - where we do not use n-grams and the linguistic features) actually helps detect more hate.

- **Disability hate:** for the disability dataset, the Comprehensive-classifier again resulted in the best overall performance with the following scores: precision (P)=0.74, Recall (R)=0.68 and F-measure (F)=0.71. It detected 68% of the hateful samples, which was the highest among the three classifiers, whereas the Othering+ raw-classifier, which resulted in precision(P)=1.00, Recall(R)=0.55 and F-measure(F)= 0.71, detected all the non-hateful samples but approximately half of the hateful samples were missed. This could be interpreted as 'othering language' being ineffective enough for detecting disability hate speech. In addition, this shows the usefulness of the (n-gram and the linguistic features) features for detecting the disability hate as the disability context was enriched by additional information obtained from the linguistic features and the different k words sequences (n-grams features), which perhaps, appears to capture the aspect of disability hate. The Comprehensive-classifier showed a lower AUC score=0.86 compared to the Othering+raw-classifier which had an AUC score=0.99, despite their F-measures being equal. However, a high AUC for the Othering+raw-classifier was expected because the classifier detected all the true negative samples ('perfect' recall)[6].

- **Racism hate:** for the race dataset, the Othering+raw-classifier resulted in precision(P)=0.84, Recall(R)=0.95 and F-measure(F)=0.92 which improved the detection of hateful and non-hateful instances compared to the Comprehensive-classifier and the Othering-classifier by 6% and 1%, respectively. Perhaps the low FNs obtained by the othering+raw-classifier means that the othering language is used more in the hate samples than the non-hateful samples in the same data. As a consequence, this made the othering+raw-classifier able to distinguish the hate from the non-hate. This suggests that the use of 'othering' features is more important than additional n-grams or linguistics features as the use of n-grams and linguistic features did not, therefore, positively contribute to the detection of racial hate speech. Also, the AUC score = 0.92 is the highest for the Othering+raw-classifier compared to the other classifiers. As mentioned previously, high AUC and high F-measures indicate that the classifier performs well at all thresholds [205].

- **Sexual orientation hate:** for the sexual orientation dataset, the Comprehensive-classifier detected the highest number of hateful instances (three were missed), with Recall(R)=0.98, suggesting that this classifier is the best at detecting hate based on sexual orientation. However, it also had the highest FPs (false detection for non-hateful instances) with Precision(P)=0.31 and F-measure(F)=0.47. The

Othering+raw-classifier missed 13% of the hateful samples, with Recall(R)=0.86, Precision(P)=0.61 and F-measure(F)=0.72, which is considered to be a more balanced classifier for the sexual-orientation dataset.

For homosexual hate, the 'othering' features alone helps balance the classifier, but having the extra features in the comprehensive classifier (n-grams and the linguistics features) actually helps detect more hate. This means that the othering feature alone is not highly effective for detecting the sexual orientation hate speech.

In summary, the Othering+raw-classifier detected the highest number of hateful samples for religion and racism while the Comprehensive-classifier performed best at detecting hateful samples in the disability and sexual orientation datasets.

This suggests that the use of 'othering language' features is more important than additional n-grams or linguistics features for detecting online anti-religion content and racism, while 'othering language' would appear to be less effective for detecting disability and sexual orientation hate. This means that different types of hate speech have various language characteristics, and the use of othering terms can be sufficient for some but not all contexts of hate speech.

4.3.2 Qualitative Results

Given the resultant improvements over the state-of-the-art using the othering feature, a qualitative analysis was conducted to identify any insights into the features captured using feature embedding on the othering features, i.e. the two-sided pronouns. Given the improvement, it can be assumed that the embedding method did effectively assign othering features to similar vector spaces in such a way as to better distinguish hateful from non-hateful content using the Paragraph2Vec embedding algorithm. Two datasets were visualised - the unprocessed Training Dataset using embeddings only (see Fig 4.8), and the othering feature set 2 with the embedding model (see Fig 4.9)

- which reflected the representation of the ‘us and them’ narrative that produced the third experiment in the last row of Table 4.6).

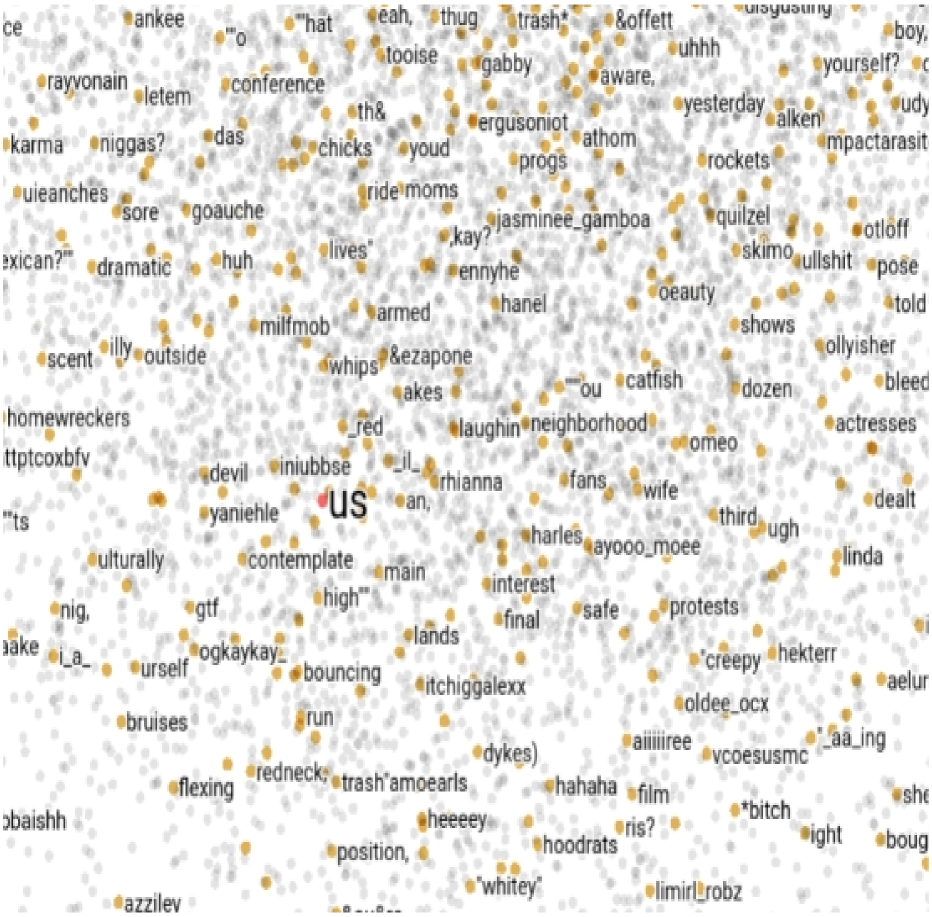


Figure 4.8: Embedding visualisation on original training data set

The model was visualised using TensorBoard which has a built-in visualiser (we performed 2D principal component analysis (PCA)), called the Embedding Projector, for interactive visualisation and analysis of high-dimensional data like embeddings [4]. The distances between words are relative based on their computed similarity to other words in the hateful sample.

The two graphs are focused and enlarged to show the 300 most similar words. The colours indicate the distances from the key word ‘us’; the purple dots indicate the smallest distances (0.008-0.09), next smallest are the pink dots (0.093-0.2), then orange

[4] https://www.tensorflow.org/versions/r0.12/how$_{t}os/embedding_{v}iz/$

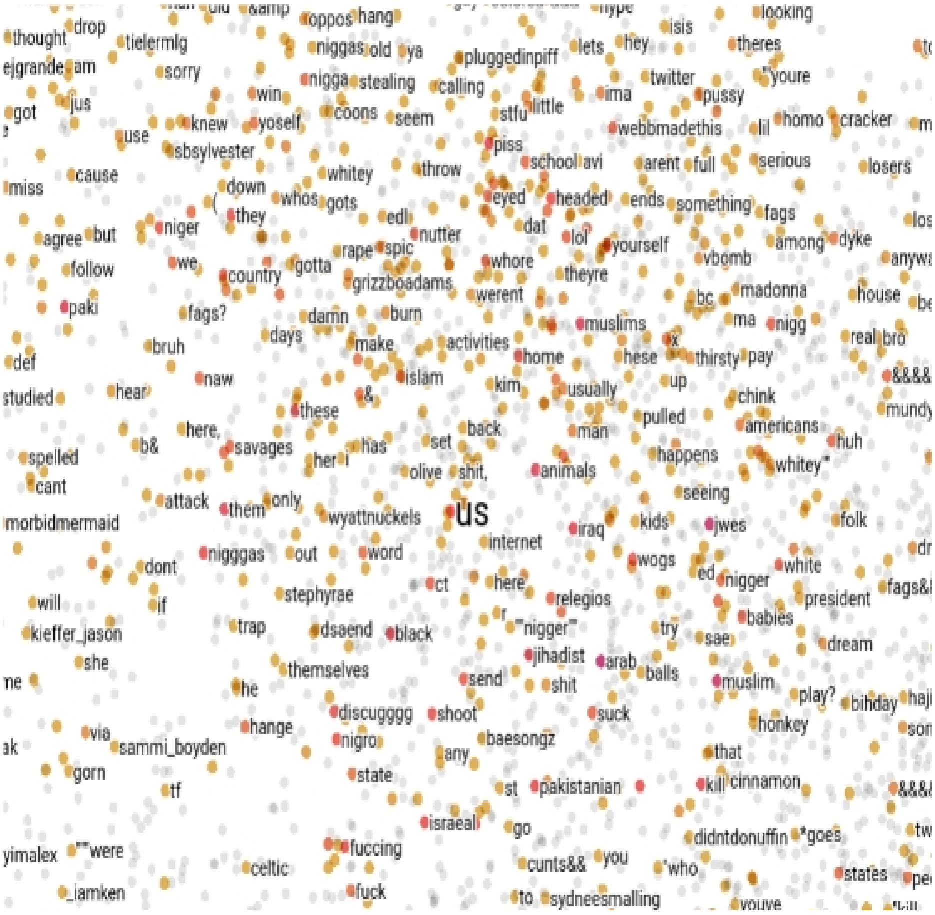


Figure 4.9: Embedding visualisation on the proposed Othering feature set 2

dots (0.21-0.39), then dark yellow dots (0.4-0.55), and finally the light yellow dots are furthest from ‘us’ (0.56-1). In distance functions, smaller values imply greater similarity between words [13].

Ideally, we want the classifier to be able to use these small distances to make effective use of them as features for distinguishing hate from non-hate.

It can be seen in Figure 4.9 that the words with the smallest relative distance from ‘us’ (the purple, pink dots) include pronouns from the ingroup (*e.g. we*), pronouns from the outgroup (*these, them etc.*), the most overt linguistic markers of alliance and distance. In actual linguistic terms, ’expressions that are most revealing of the boundaries separating Self and Other are inclusive and exclusive pronouns and possessives such as *we* and *they*, *us* and *them*, and *ours* and *theirs*' [139]. These linguistic items became semantically similar because they were used in the same context when the writers were trying to distance themselves from the other. In addition, different action verbs (*send, go etc.*) appear, which capture their co-occurrence in more nuanced othering aspects of cyberhate language. From an Integrated Threat Theory perspective, we can also see symbolic and realistic anxiety present (e.g. the words 'attack', 'state', 'jihadist', 'animals') and intergroup anxiety when there are feelings of discomfort that people may experience when engaging with members from a group other than their own, which can also be referred to as the 'anxiety that people experience in interactions with members of another group' [47] (e.g. 'Arab', 'Israel').

Furthermore, we can see symbolic threats which are concerned with a group's values, traditions, ideology, morals, and these are expected to be more prominent when an in-group believes that their cultural values and traits (e.g. appearance) are different from those of an out-group (e.g. Muslims, Jews, Arab, Pakistani, Africano, American). We can also see the obvious derogatory terms in the same Figure (e.g. *suck, fuck, discuggg, niger, niggero, savages*). Thus, this model is picking up both the obvious and non-obvious cyberhate using the proposed othering features.

Meanwhile, in Figure 4.8 none of the words are particularly close to 'us' in terms of distance, meaning the classifier is unable to make effective use of the othering narrative. Thus, the classifier based on these features will become more dependent on the hateful words and miss the less obvious narrative. We posit that this work demonstrates that the ability to capture the more nuanced text is the core reason behind the successful improvement over the state-of-the-art examples from previous research. In Table 4.8, the visualisation is summarised by showing the top 10 similar words through the two models: embedding based on the unprocessed dataset (raw training dataset) only, and the othering feature embedding. The similarity was measured by using the cosine similarity function. Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space [154]. The similarity table shows the most similar words to the word *‘us’* using the model which reflects the definition of different groups (e.g. *’Muslims, Jews, jihadist’*) while the word *’us’* in the raw embedding refers to concepts (e.g.*’fans, safe, devil, wife’*). The othering feature set succeeds in defining different groups and different attitudes which are important aspects in the field of hate speech recognition.

Table 4.8: Target words and their 10 most similar words as induced by unprocessed dataset (raw training data set) embedding and the proposed 'othering feature set 2' embedding.

Target word	Raw training Data Set Embedding	Othering Feature Set Embedding
us	fans, safe, devil, wife, gisuz, whips main, they, tinge, armi	ni**as, we, Arab, those, iran group, Jews, Muslim, send, headiest,animals
them	sort, stop, close, ask, differ, speak busi, tweeting, ni**ers, we	these, Jew, those, pakistanian someon hang, Muslims, rednecks, country, p**s, niggu
arab	shit, hot, like, bleiv, thing hat, went, thing, die, lol	iraq, we, ni**a, fu**er, israel, getout nig******as, outta, chop, animals
Muslim	smh, office, appear, backlash, nonmuslim dye, agenda, rasicm, christian, asian	us, nonmuslim, iraq, lable, Jews, racism high, arab, yellow, doe, country
send	only, what, neireian, fu**ing, I bet, saying, realy, reason, pick	Israel, suck, Africano, ni**er, getout, home, these paki, ni**o, kill, burn
out	have, try, see, never, some, home bastards, fail, wrong, give	gotta, outta, Islam, America, chop, attack, mosque, send home, manchest, blacks, these
scum	hear, speak, dumb, edltrobinson, exact, sort, breedingwoolwich, seem, kind, threaten	savages,islamic, yell, muslims, beat, pi**, tuesday wogs, against, burn, qouet, leage
shoot	divis, behind, pub, bloodi, scence hospital, condemen, plane	fu**ing, shot, nobody, ni**o, disable nutter, muslims, burn, nonmuslim, condemen
kill	stabb, death, pari, involve, year between, brutal, cheldren, charge, three	stabb, live, them, innocent, pakistanian soldir, who, place, krazi, suspect
burn	local, critic, swedish, bbcnews, islamabad faggots, saudi, antiterror, bangladish, letter	church, non-whites, chop, mosques, themfu**ing,shoot chop, ni**as, commit, nigeria

4.4 Conclusion

This chapter has shown how this research aimed to improve the machine classification performance for different types of hateful and antagonistic language posted on Twitter - known as cyberhate. This study was inspired by the concepts presented by Integrated Threat Theory (ITT) and ‘othering’ theory.

The hypothesis was that the use of an ’othering feature set’ would provide better context for the classifier beyond using words alone. Vector embedding and the Paragraph2Vec algorithm were used to cluster these features, thereby re-framing the linguistic features from individual terms and phrases to numeric distances representing a form of semantic similarity of these terms - learned in the context of hateful or non-hateful texts. Then, machine classification methods were experimented with to determine the improvement of the novel ‘othering feature’ over the state-of-the-art research, and the most effective machine classifier to use with the embedding-transformed feature set.

Generally, the results show the effectiveness of including the proposed feature set for classifier training over the baselines model and improved the rate of the performance of the baselines’ classifiers in the cyberhate literature by 2-59%, showing the ability of the ’othering’ narrative to be an important part of hate speech detection. When tested on completely different datasets using four different types of cyberhate, namely religion, disability, race and sexual orientation, F-measures of 0.81, 0.71, 0.89 and 0.72 respectively were obtained across two models; however, the models perform well on some but not all categories of the unseen data (types of hate speech).

This indicates that different types of hate speech have different language characteristics and the use of othering terms can be effective for some but not all contexts of hate speech. The experiments suggest that othering language can indeed be a valuable feature for identifying anti-religious and racial content while it is less valuable as a standalone feature for disability and sexual orientation hate.

Naturally, the approach taken in this work involves some limitations; first, a dataset with a considerably small sample size was used for training and testing this thesis classifier. It is preferable to implement embedding learning on a large amount of text data to ensure that valuable embeddings are learned. This limitation was due to the limited number of relevant datasets that are publicly shared and also the high price of human annotation. Second, by testing the proposed model, it was assumed that tweets which contain two-sided othering language (othering pattern) are more likely to be hateful. However, this is not necessarily the case. For example, the tweet (e.g. "send them") may imply hate if it appears in relation to a hateful event, but neutral when it appears in relation to a marketing-related tweet. The point here is that the implicit cyberhate needs more studies to investigate its patterns.

In the end, the contextual hate, discussed in this chapter, that interested the author is posted by individuals and groups who have several audiences - these may include the victimised target group of the hate communication, ideological friends and potentially interested groups of persons that may be recruited, as well as a more general audience. The question here is who are these groups, how are they connected and how do they communicate? The next chapter discusses the investigation of a study of social dynamics in online social networks in terms of (i) individuals' exposure to online hate, and (ii) individuals' roles in propagating online hate among groups - more specifically, developing insights through Social Network Analysis (SNA) into multiple hateful networks with the aim of characterising and understanding these networks.

Chapter 5

Networks of Hate

5.1 Introduction

The previous chapter focused on detecting cyberhate automatically, a method which offers the ability to create representations of how online hate is composed and communicated. As online social media enables individuals and groups to spread ideologies and even advocate hate crime, it is essential to study the online structure, connectivity and communication of online communities in order to determine users' exposure to and the propagation of hateful ideologies that could influence their own views and actions.

Individuals and groups have increasingly used the internet to express their ideas, spread their beliefs, and recruit new members [193]. As with offline hate crime, cyberhate posted on social media has become a growing social problem. In 2016 and 2017, the UK's decision to leave the European Union and a string of terror attacks were followed by noticeable and unprecedented increases in cyberhate [332]; the rhetoric was that of invasion, threat and otherness [27]. Some research suggests that the perpetrators of cyberhate have similar motivations to those who resort to violence offline [332, 33, 72, 34]. Social psychologists have suggested that the perpetrators of hate crime may be influenced by their perception that certain groups pose a threat to them [298], and Glaser *et al.* [133] suggests that racists often express their views more freely on the internet than elsewhere. Research reveals that exposure may be correlated with detrimental effects on a societal level, as exposure is potentially linked to an increase in hate crimes offline, a lack of social trust, tougher discrimination, and prejudice against the targets[175, 236]. This includes spreading extremist and violent ideology[120, 312]. Twitter, which has become an essential source of up-to-the-minute information, offers a unique opportunity to study social dynamics in online social networks in terms of (i) individuals' exposure to online hate, and (ii) the role of individuals in propagating online hate among groups.

The detection of hate online has been widely discussed from the perspective of content analysis. Furthermore, the previous chapter has shown an improvement in automating the detection of hateful content. However, the study of hateful networks on social media has received limited attention in the literature. Given we can now detect hate online automatically using machine learning methods, a study of such networks could be valuable in the context of concern about the connectivity among hateful users and therefore the exposure to, and contagion of, online hateful and offensive narratives on social media. On the Twitter platform, the hateful *followers' network* represents the hateful user community directly exposed to hateful content. This network is a subset of users who directly receive information from each other. Furthermore, the hateful *retweets' network* is a construct formed by users who propagate cyberhate to their own followers, thereby passing on hateful narratives from the people they follow - a form of cyberhate contagion.

It is important to note that people who are exposed to hateful content won't necessarily spread hate. However, the exposure to hateful content among the hateful users refers to the existence of communities interested in this topic which potentially increases the risk of more people adopting hateful ideologies.

Several studies have applied Social Network Analysis (SNA) to Twitter hateful networks in order to use connectivity information as an indicator that a user is posting offensive content [273, 19]. Others have focused SNA analysis on the retweets network to measure diffusion [278, 273]. However, there is yet to be a study of *multiple* hateful networks with the aim of understanding whether there is evidence of similar ’levels of friendship’ among hateful users and therefore a general connectivity and exposure to the hate, or similar levels of propagation behaviour among hateful users and therefore a general contagion effect. The lack of such a study on Twitter motivated the author of this research to undertake a baseline study that characterises several hateful networks extensively from multiple perspectives, namely: exposure to cyberhate (in followers’ networks) and propagation of the cyberhate (in retweets’ networks). Investigating such characteristics is a basis for a comprehensive understanding of online hateful networks, which could help decision-makers to mitigate the danger of these networks. Note that a network is labelled as a hateful network if all users belonging to that network have posted tweets that human annotators agree should be classified as containing evidence of hateful content.

The remainder of the chapter is organised as follows. Section 5.2 describes the methods of the present research, including data collection. Section 5.3 reports the results and the discussion and Section 5.4 presents the conclusions .

5.2 Methods

This section is a reflection of the design and development phase mentioned in Chapter 3 by designing multiple hateful networks in order to understand the characteristics of hateful online social networks in a small non-representative sample. Also, this section reflects the demonstration phase that has been mentioned in Chapter 3 by applying a range of network analysis metrics for characterising several hateful networks.

5.2.1 Datasets Build

Figure 5.1 illustrates the steps for collecting and building Anti-Muslim 1 and Anti-Muslim 2 datasets for this study.

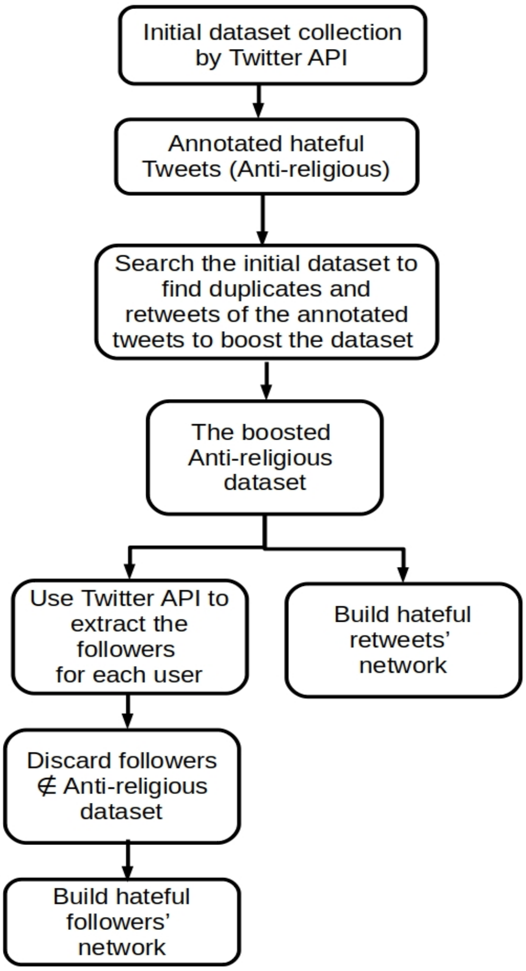

Initial dataset collection by Twitter API
Annotated hateful Tweets (Anti-religious)
Search the initial dataset to find duplicates and retweets of the annotated tweets to boost the dataset
The boosted Anti-religious dataset
Use Twitter API to extract the followers for each user
Build hateful retweets' network
Discard followers ∉ Anti-religious dataset
Build hateful followers' network

Figure 5.1: Data collection and build steps for the hateful followers' and retweets' networks.

To remind the reader, the collection for Anti-Muslim 1 and Anti-Muslim 2 resulted in 427,330 tweets and 919,854 tweets, respectively. Initially, to identify hateful instances in these datasets, we applied the 'Othering+raw classifier' from the previous chapter to classify a sub-sample of the datasets. To validate the outcomes, we sent the hateful cases for human annotation. We chose the 'Othering+raw classifier' because it resulted in detecting the highest number of anti-religious instances (see Chapter 4, Table4.7). However, implementing the 'Othering+raw classifier' requires the extraction of Typed Dependency features for the tested dataset (see Figure 4.7 - Chapter 4). Extracting Typed Dependency features for a large text corpus required more computing resources than were available to the author, which forced the author to skip this particular pre-processing step. Therefore, a reduced feature-set was implemented for the 'Othering+raw classifier'. While this may reduce performance (i.e. miss some hateful instances and provide less hateful content for this study), a human annotation step was included to ensure the outcomes labelled hateful by the classifier were accurate (i.e. were actually hateful).

Due to the available resources, this classifier was applied to two subsets each containing 40,000 tweets, taken from each 10 tweets of each initial dataset. This resulted in 16531 (41%) and 12032 (30%) being classified as hateful for Anti-Muslim 1 and Anti-Muslim 2, respectively. Two subsets, each of 4000 tweets taken from each 10 tweets of the resulting dataset, were chosen for a human annotation process. Human annotators were asked to label the offensive tweets using the crowd-sourced online service Crowdflower. Annotators were provided with each tweet, and asked 'Is this text offensive or antagonistic in terms of race, ethnicity or religion?' They were presented with a ternary set of classes: yes, no, undecided. The results from coders could then either be accepted or rejected on the basis of the level of agreement with other coders. The requirement was that at least four human annotations per tweet and only the annotated tweets for which at least three human annotators (75%) had agreed on the appropriate class were retained, as per related work [308, 60]. Annotators were hired with no special background knowledge, who were aged 18 and above. CrowdFlower makes it possible to require workers to come from English-speaking countries, a feature that other platforms like Amazon Mechanical Turk do not offer transparently. It has a built-in quality control mechanism, using interspersed test items, ensuring that workers maintain a certain level of accuracy throughout the entire job. The annotation process took about 5 to 9 days to complete. [1].

[1] An ethical approval for the data collection has been included in the Appendix D. The wording of the research question has been revised slightly as the thesis has evolved and been through peer review,

The results of the annotation exercise produced 'gold standard' datasets with 973 of 4000 and 1053 of 4000 instances of offensive or antagonistic content tweets for the Anti-Muslim 1 and Anti-Muslim 2 datasets respectively.

The interest in these data was to flag the Twitter accounts of users posting hateful content. The initial datasets, which were collected by Twitter API at the stage of the data collection, were searched to uncover any duplicates of the annotated hateful tweets (tweets with the same text posted by other users). This boosted the collection of hateful tweets to 2621 and 2097 tweets for Anti-Muslim 1 and Anti-Muslim 2, respectively. Finally, we extracted the accounts of those who retweeted the content, creating a list of 3502 and 8602 user accounts that were involved in creating or propagating hateful content for Anti-Muslim 1 and Anti-Muslim 2 respectively. Note that we collected all accounts that posted hateful content, whether the tweet was the original post, a duplicate or retweet. The retweets were extracted using a pattern recognition technique to extract any post matching the following format: 'RT '+ space + '@screenname' + space + ':' + 'Tweet text' [88]. The pattern shows that the collected retweets do not contain a 'tweet quote'. This step was taken because the hateful retweeters may 'quote retweeted' the tweet ironically or to express the rejection of the tweet's content.

From each dataset (Anti-Muslim 1 and Anti-Muslim 2 datasets), it was necessary to extract the hateful followers' dataset (hateful users who follow each other) and a retweet dataset (users who retweet each other). For the followers' datasets, for each of the authors of the 3502 and 8602 tweets classified as containing evidence of possible hateful speech, a list of followers was retrieved for these accounts so that the measures of connectivity between these users could be identified. This collection resulted in two sets of 2,018,950 and 3,855,37 followers for lists of 3502 and 8602 authors, respectively. Note that the users in these sets (2,018,950 and 3,855,37) were not necessarily expected to be hateful users.

but they are substantively the same in principle as when they were captured in the ethics paperwork, and the use of data has not changed in purpose since approval

Next, Python tools[2], Panda package were used to generate two types of networks of followers and retweets - (see also Section 5.2.3). The followers' network was built using the followers' dataset. This is a directed graph network in which each node has a follower $h \xleftarrow{follow} x$ relationship.

Algorithm 5.1 explains the steps for building a hateful followers' network. It shows that any follower relationship of users who had not been shown to post hateful content was discarded, so that we could identify measures of connectivity between only hateful users. Overall, 1004 users and their 2644 followers were extracted from the Anti-Muslim 1 dataset, and 1073 and their 2895 followers were extracted from the Anti-Muslim 2 dataset. This led to two datasets of followers that contained the original users and their followers (those exposed to cyberhate and those who had also posted cyberhate) - one for Anti-Muslim 1 followers and the other for Anti-Muslim 2 followers.

[2] https://www.python.org/

Algorithm 5.1 Building the hateful Follower Network

```
INPUT: $H = h_1, h_2, .., h_n$users accounts who post hateful content
OUTPUT: hateful Follower Network $H_{followers}$
for each $h_i \in$ H do {do until the end of the list H}
    Collect follower network $h_{fo}$
    for each follow relation $h_i \overset{follow}{\longleftarrow} x_i \in h_{fo}$ do
        if $x_i \in$ H then
            Return adjacency list $H_{followers}[h_i,x_i]$
        else
            discard the relation $h_i \overset{follow}{\longleftarrow} x_i$
    end if
    end for
end for
Return adjacency list $H_{followers}$
```

For the retweets datasets, two retweets' networks were built: Anti-Muslim 1 with 1229 nodes and 2571 edges, and Anti-Muslim 2 with 5581 nodes and 16338 edges. Each of these is a directed network having a i and $j \in$ retweets' dataset for each node edged from i to j, indicating that j is a retweeter of a tweet posted by i. Algorithm 5.2 explains the steps taken to build a hateful retweet network. It shows that we created the retweets' networks from the 3502 and 8602 user accounts that were involved in creating or propagating hateful content for Anti-Muslim 1 and Anti-Muslim 2, respectively. Note that not all hateful accounts are retweeters.

Algorithm 5.2 Building the hateful retweet Network

```
INPUT: $H = h_1, h_2, .., h_n$ users accounts who posted hateful content
OUTPUT: hateful retweets'network $H_{retweet}$
for each $h_i$, $h_j \in$ H do {do until the end of the list H}
    if $h_j\ retweeted\ h_i$ then
        Return adjacency list $H_{retweeters}[h_i,h_j]$
    else
        Return null
end if
end for
Return adjacency list $H_{retweeters}$
```

Figures 5.2 shows a sample visualisation of the retweet network for the Anti-Muslim 1 dataset.

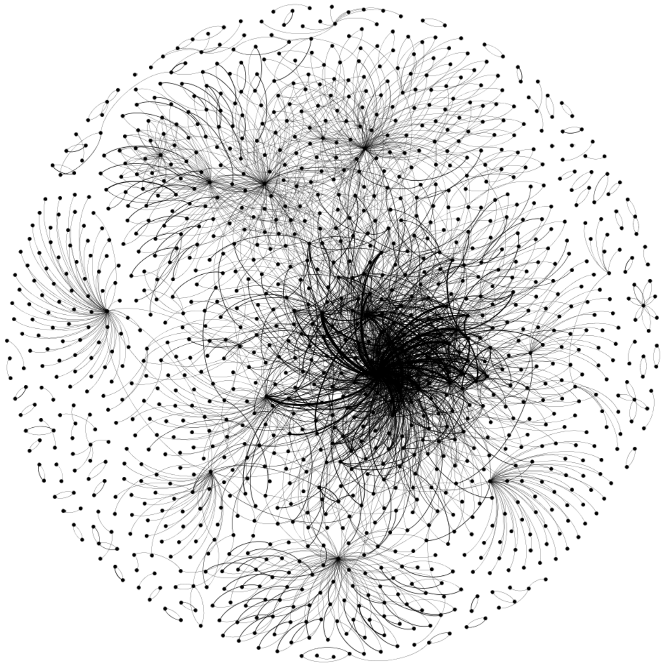


Figure 5.2: Example of a graph representation of the retweets graph of Anti-Muslim 1 users.

5.2.2 Human Annotation vs. machine classification performance

This section evaluates the quality of 'Othering+raw classifier', which performed the best in terms of detecting hateful religious tweets and 'Comprehensive-classifier', which produced a more balanced classification (but higher FNs than 'Othering+raw classifier'). The evaluation compares the outcomes of the human annotation with the classifier results. Table 5.1 shows the results of applying our classifier to the annotated datasets: Anti-Muslim 1 with (*973 hateful instances and 4000-973= 3027 non-hateful instances*) and Anti-Muslim 2 with (*1053 hateful instances and 4000-1053=2947 non-hateful instances).*

This indicates that the *'Othering+raw classifier'* detected 78% and 74% of the hateful samples for Anti-Muslim 1 and Anti-Muslim 2, respectively. Additionally, *Comprehensive-classifier* detected 69% and 71% of the hateful samples for Anti-Muslim 1 and Anti-

Table 5.1: Machine classification performance for *’Othering+raw classifier’*and *Comprehensive-classifier* based on unseen testing datasets: Anti-Muslim 1 and Anti-Muslim 2 .

	Anti-Muslim 1			**Anti-Muslim 2**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Othering+raw classifier	0.91 FN=207	0.71 FP=748	0.82	0.90 FN=211	0.69 FP=899	0.78
Comprehensive-classifier	0.89 FN=297	0.83 FP=494	0.86	0.88 FN=302	0.82 FP=528	0.85

Muslim 2, respectively. This suggests that our model is less accurate at detecting religious hate than the human annotation by 29-31%; this is a ’solid’ performance (percentage of detection is about 70% or above)[279]. Interestingly, the two classifiers showed similar behaviour to their performance on the religious datasets in Chapter 4, *when the ’Othering+raw classifier’* achieved lower FNs while *Comprehensive-classifier* appeared to be a more balanced classifier but involved a higher risk of missing hateful comments (FNs). The similarity of the models’ behaviours on different unseen datasets may lead to the assumption that the test datasets represent the distribution of future cases of anti-religious content. In addition, the ’solid’ performances, of 69%-86%, indicate the usefulness of the ’othering feature set’ for detecting the hateful content, perhaps the implicit hateful content which is one of the main obstacles to detecting online hate [350]. However, we still have high FPs (high neutral tweets classified incorrectly as hateful tweets), which means that our feature sets need more work in order to make the classifier more balanced. This also suggests that our classifier could be used to improve the annotated dataset results. By way of example, the human annotation for Anti-Muslim 1 resulted in 973 hateful instances out of 4,000 instances, which is only 24% of the annotated instances. This is one reason for the scarcity of datasets in this field. The annotation is expensive, and the results achieved for annotated hateful samples usually appear small in size. This may be because of the method for choosing sample tweets to send to the annotators. The aim here is thus to increase this percentage. It may be possible to achieve this by applying our classifier first to the whole original dataset and then gathering a sample of tweets classified as hateful and sending those to human annotators. In fact, this suggestion requires further experimentation regarding text analysis and classification. For this research, we depended on the results of human annotation that was undertaken first (resulting in 24% and 26% of hateful instances for Anti-Muslim 1 and Anti-Muslim 2, respectively).

Anti-Semitic Dataset

As with the Anti-Muslim data, the annotated Anti-Semitic dataset was extended by adding duplicates and retweets of the original tweets from the larger data collection. This boosted the annotated dataset to 3874 tweets. For the Anti-Semitism dataset, a followers' network was not extracted because Twitter API did not recognise all the relevant users' IDs. However, the retweet network for Anti-Semitism, which consisted of 2748 nodes and 5091 edges was built.

Dataset Implications

The above datasets were collected at different time and in different circumstances. While the anti-Muslim datasets were collected in two weeks, the antisemitism dataset was collected over one year. Furthermore, the Anti-Muslim 1 dataset was collected from Twitter for a period immediately following the Woolwich trigger event, containing particular characteristics in terms of religion and race. This reflects the emotion following large scale, disruptive, and emotive events such as terrorist attacks in near real-time. Anti-Muslim 2 is broader than Anti-Muslim 1 as it was collected in the period of publication of a letter that promoted hate toward a specific social group, reflecting more general responses and expressions in term of religion toward a specific social group. The antisemitism dataset was collected using antisemitic keywords without any trigger event, which is expected to contain broader hate toward a specific social group than Anti-Muslim 1 and Anti-Muslim 2. Based on the differences, we argue these factors suggest that any similarities between these datasets (networks) might be considered common characteristics of hateful groups.

Comparative "Risky" Network Dataset

Although Twitter networks of a different size and nature inevitably show different characteristics, the larger the number of common properties (metrics), the more likely it is that the two networks are similar[267, 225]. In this work, the term 'risky networks' was defined as those that might include dangerous/unsafe content. Comparing a non-hateful (but risky) network with hateful networks may help to understand the level of the similarity and the differences among the hateful networks themselves. For example, networks A and B with sizes 1 and 2 may not be similar. However, comparing networks A and B with another network C with size 6 shows that networks A and B are slightly similar and different from network C. In addition, we assume that if the hateful networks' connectivity level is higher than another risky network, they exhibit more dangerous consequences (e.g. propagation), increasing the exposure to at-risk users. Thus, for the purposes of comparison between networks in this study (e.g. do hateful networks exhibit different characteristics to other risky networks?), a similar size network from another 'risky' category was selected - one in which users in online social networks risk exposure to ideology and where there is concern about the contagion of content. It was found that the suicidal network by Colombo *et al.* [88] was similar to our networks in terms of three factors: (1) comparable size, (2) similar data collection process (Twitter API) and network build, and (3) likely to spread content of concerning ideology, i.e. a 'risky network'. Table 5.2 summarises the number of nodes and edges for each network.

Table 5.2: Numbers of the nodes and the edges of the hateful networks

Followers’ networks		
Network	**Nodes**	**Edges**
Anti-Muslim 1	1004	2644
Anti-Muslim 2	1073	2895
Suicidal	987	2410
Retweets’ networks		
Network	**Nodes**	**Edges**
Anti-Muslim 1	1229	2571
Anti-Muslim 2	5581	16338
Anti-Semitic	2748	5091
Suicidal	3209	2211

As with the hateful datasets, the suicidal network has two sub networks - the followers and retweet networks. Both networks are of a similar size to the networks examined in this study. The suicidal content was also labelled by human annotators in the original paper. For the suicidal network, the followers’ network contains 987 nodes and 2410 edges, whereas the retweet network contains 3209 nodes and 2211 edges.

5.2.3 Metrics Selection

Social network analysis relies heavily on quantitative features to numerically define various attributes of the network. These featuresare also referred to as social network metrics. Value of various SNA metrics discussed below can be used to answer questions that help to understand the structure of a network, the flow of information in the network and important individuals in the network. In Chapter 2, the meaning of social network analysis is explained. A graph is composed of two fundamental units: vertices (also called nodes) and edges. Every edge is defined by a pair of nodes. In this chapter, nodes represent the hateful author and, in turn, an edge is a line that connects two nodes and, analogously, represents either a **’follow’** or **’retweet’** relationship. Edges may be directed or undirected; in this study, all edges are directed. In a directed graph G, with a total number of edges $E(G)| = m$, the maximum number of edges $m_{max} = n(n-1)$.

Social networks graphs were built using the followers and retweets datasets. Then, the metrics were extracted and compared. As discussed in [267, 225], the larger the number of common properties (metrics), the more likely it is that the two networks are similar. For the implementation of the graphs' metrics, we used NetworkX. NetworkX is a package for the Python programming language. The implementation of NetworkX is advantageous due to its speed of implementation. Gephi was also partially used. Gephi is an open-source network analysis and visualisation software package written in Java on the NetBeans platform. We used Gephi for file exporter (File > Export > Graph file...) to save the hateful network into a file that could be imported by NetworkX. Additionally, Gephi was used for network visualisation. The metrics used in this study were selected and implemented as follows:

- **Giant component:** The Giant Component (GC) is a connected component of a given graph that contains a finite fraction of the entire graph's nodes, e.g., a significant proportion of the nodes are connected in one GC. The GC of the networks was extracted using a Depth-first Search and Linear graph algorithms [307]. Formally, let $s(n)$ be the size of a connected component GC in a network of size N, then GC is a giant component if:

$$\lim_{n \to \infty} \frac{s(n)}{N} = N > 0 \tag{5.1}$$

  where n is the number of nodes. This limit would go to zero for all other non-giant components. From a hate spread perspective, the size of the GC is essential in that it reveals the maximum number of people who can be (directly or indirectly) reached by any other node in the same component. A large GC indicates high reachability because every node is reachable from almost every other. Networkx implementation of the GC is as follow:

```
import networkx as nx
```

```
giant = max(nx.connected_component_subgraphs(G), key=
    len)
```

- **Density:** The ratio between the number of edges in the graph and the total number of possible edges, as defined in Equation 4.1.

$$Density = \frac{m(G)}{m_{max}(G)}, 0 < density < 1 \tag{5.2}$$

where m(G) is the number of edges in the network and m_{max} *(G)* denotes the number of possible edges, which is *n(n - 1)* Density measures how close the network is to completion. A complete graph has all possible edges and density equal to 1. The opposite, a graph with only a few edges, is a sparse graph [360]. High density indicates intimate, tightly knit networks, and ties between individuals in a denser network are more likely to be strong. Networkx implementation of the networks density is as follows:

```
density=nx.density(g)
```

- **Average degree metrics:** are direct measures of how information travels throughout the network [239]. Average graph degree for a node is calculated as the number of links that end in that node. Also, the maximum value of the degree of the nodes over all the graph nodes was calculated, as defined in Equation 4.2.

$$AverageDegree = \frac{TotalEdges}{TotalNode} \tag{5.3}$$

Essentially, this metric is a measure of graph connectivity in terms of links/relations between nodes. This, in terms of followers' degrees, means that users can directly consume (see, read) the content posted by other users. The spread of nodes' degrees over a network is characterised by a distribution function, which is the probability that a randomly selected node has exactly k edges. The degree distribution has been calculated for the followers and retweet networks.

In this study, the research interest lies in the out-degree, which represents the number of users that someone follows (e.g. if A has an out-degree of 5, it means A follows five people). Also, the in-degree distribution is calculated, which represents the number of followers that someone has (e.g. if A has an in-degree of 5, it means there are five people that follow A). Higher out-degree values means a wider exposure to different sources of hate propagators. Higher in-degree value refers to influential users (content creation hubs or conversational hubs) who can be responsible for hate creation and propagation.

For the retweet network, out-degree represents the number of retweets (e.g. if A has an out-degree of 5, it means they retweeted five tweets posted by five different users). Also, we are interested in the in-degree distribution that shows the number of retweets that someone gained (e.g. if A has an in-degree of 5, it means five users retweeted A's tweet). A high in-degree indicates high hate propagation, while a high out-degree indicates a the level of diversity in the propagated content. Networkx implementation of the networks' average degree as follow:

```
AverageDegree= nx.average_degree(G)
```

- **Average clustering coefficient:** Firstly we calculate the clustering coefficient for each node as the probability that two randomly chosen distinct neighbours of the given node are connected; this is also referred to as the local clustering coefficient for a node. This coefficient is, therefore, given by the fraction of pairs of nodes, which are neighbours of a given node that are connected to each other by edges. See the following equation:

$$c_i = \frac{2\,|e_{jk}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E \tag{5.4}$$

where N_i is the neighbourhood of node v_i , e_{jk} represents the edge that connects node v_j to node v_k , k_i is the degree of node v_i , and |e_{jk} | indicates the proportion of links between the nodes within the neighbourhood of node v_i. Then these values are averaged, all local values c_i over all the network nodes. The average clustering coefficient (C) was calculated using the Matthieu Latapy algorithm [189]. See the following equation.

$$C = \frac{1}{n\sum_i} c_i \tag{5.5}$$

The clustering coefficient measures how some of the nodes can form dense groups in which each element has strong connections with the others. As a consequence, each piece of information posted by one of these nodes can rapidly spread within the groups but disseminates outside the group with more difficulty. Networkx implementation of the CC is as follow:

```
AverageClusteringCoeffecient= nx.average_clustering(G
    )
```

- **Reciprocity:** Reciprocity r is a specific quantity for directed networks that measures the likelihood of nodes in a directed network being mutually linked. See the following equation:

  $$r = \frac{\#mut}{\#mut + \#asym}, 0 < r < 1 \tag{5.6}$$

  where mut denotes the number of mutual dyads and asym the number of asymmetric dyads (an asymmetric dyad is a pair of nodes that has an arc going in the direction of one node or the other, but not both directions). For two nodes i and j in a given graph G and the related adjacency metric A, if the relationships $A(i, j)$ and $A(j, i)$ exist, then this is considered a reciprocated relationship. In the case of a follower network, if $i \overset{follow}{\longleftarrow} j$ and $j \overset{follow}{\longleftarrow} i$, then they have a reciprocated follow relationship. For the retweet network, if $i \overset{retweet}{\longleftarrow} j$ and $j \overset{retweet}{\longleftarrow} i$, then they have a reciprocated retweet relationship. A higher value indicates many nodes have two-way links, reflecting high connectivity (a high level of friendship) in the followers' network and high cooperation for hate dissemination in retweet networks. Networkx implementation of the networks reciprocities is as follow:

```
Reciprocity= nx.overall_reciprocity(G)
```

- **The average shortest path and diameter:** is the average graph-distance between all the pairs of nodes. See the following equation:

$$Avg = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} d(i,j) \tag{5.7}$$

where *d(i , j)* is the geodesic distance between nodes *i* and *j* , and *1/2 n(n-1)* is the number of possible edges in a network comprising n nodes. This metric gives an idea of how far apart nodes will be on average. The diameter is the longest graph distance between any two nodes in the network [22]. The Faster Algorithm for closeness centrality was used to extract the average shortest paths and the diameters [57]. These metrics were chosen because they are direct measures of how information travels throughout the network. Followers' paths represent links between a node and its neighbours, between them and their own networks, and so on. The shorter the length of the shortest path from a node to all others in the graph (and so their average), the easier the information can travel from a given node and spread over the network[88]. Networkx implementation of the ASP is as follow:

```
G= nx.path_graph(5)
 print(nx.average_shortest_path_length(G))
```

5.3 Results and Discussion

This section is reflecting the evaluation step of the Data Science Research Methodology (DSRM) motioned in Chapter 3. The evaluation step revolved around comparing and contrasting different hateful networks characteristics and comparing a ’risky’ network - characterised by language related to suicidal ideation. The evaluation process here examines if hateful networks are similar in terms of online hate exposure and online hate propagation, and more connected compared to another ’risky’ network. Tables 5.3 and 5.4 show the graph metrics for the hateful followers’ networks (Anti-Muslim 1 and Anti-Muslim 2), the hateful retweets networks (Anti-Muslim 1, Anti-Muslim 2 and Anti-Semitic), and the comparator suicidal network.

Table 5.3: Graph metrics for the followers networks

Networks	||**Nodes**||	||**Edges**||	**Giant Component**	**Density**	**Avg. Deg.**	**Max. Deg.**	Avg **Clust.**	**Avg. sh.**	**Diameter**	**reciprocity**
Anti-Muslim 1	1004	2644	60.7% (edges=66%)	0.0026 (GC=0.0047)	2.6	100	0.062	5.4	16	33.4%
Anti-Muslim 2	1073	2895	66% (edges=83%)	0.0025 (GC=0.0048)	2.7	143	0.065	5.6	17	26.7%
Suicidal	987	2410	50% (edges=45%)	0.0024(GC=0.0044)	2.53	100	0.064	5.6	17	61%

Table 5.4: Graph metrics for the retweet networks

Networks	||**Nodes**||	||**Edges**||	**Giant Component**	**Density**	**Avg. Deg.**	**Max. Deg.**	**Avg Clust.**	**Avg. sh.**	**Diameter**	**reciprocity**
Anti-Muslim 1	1229	2571	69.2% (edges=71%)	0.0017 (GC=0.002)	2.09	304	0.0097	5.2	21	18.89%
Anti-Muslim 2	5581	16338	81.3% (edges=72%)	0.00054 (GC=0.00064)	2.3	1034	0.15	5.9	16	15.61%
Anti-Semitic	2748	5091	72.1% (edges=68%)	0.00067 (GC=0.00072)	1.9	522	0.029	6.3	14	12%
Suicidal	3209	2211	31.3%(edges=24%)	0.00021 (GC=0.0005)	1.4	44	9.4E-03	5.05	13	0.9%

5.3.1 Follower Graph: measure of hateful content exposure

- **Giant Component:**

  Table 5.3 shows that the hateful followers’ networks have a similar sized GC (60.7% and 66%); the size of the giant component represents over half of the entire networks’ nodes, while the suicidal network with a similar number of nodes and edges is smaller at 50%. The size of the GC is the maximum number of people who can be exposed to/propagate hateful content. This suggests, in this small sample, that the users in hateful networks are at similar levels of risk to exposure, with suicide networks as a comparator risky network around 10% lower. Further research with much larger representative samples of networks is required to identify if this pattern is generalisable beyond this study.

- **Hate Density:**

  Table 5.3 shows that hateful networks, in this sample, have a similar and slightly higher density than the suicidal network by 0.0001. Also, the table shows that the GCs of the hateful followers' networks are similar and also slightly denser than the suicidal followers' network by 0.0002. Despite the seemingly small numerical difference, this has an impact on the rate of information flow within the network. This is because, given that hateful networks have more nodes in the GC (higher number of users) than the suicidal one, it would be expected that they would have smaller density, as in networks representing real systems (e.g. Twitter networks), the density of a network is inversely proportional to the size of that network [114][161]. Actually, they show slightly *higher* density values, suggesting that users in the hateful network are slightly more interconnected and closer to each other than users in the suicidal network. Highly interconnected users in a followers' network means increased potential for content exposure, which in turn increases the risk of potential content propagation. This suggests that lowering the density of the hateful community would solve the problem of hate exposure. However, additional research with much larger representative samples of networks is required to identify if this pattern is generalisable beyond this study.

- **Average Degree:**

  Table 5.3 shows that the hateful followers' networks exhibited 2.6 and 2.7 average degrees respectively. The hateful networks have a slightly higher average degree than the suicidal network, which is of a comparable size, while the max degree was slightly higher for the Anti-Muslim 2 network. Generally, the expected average degree of a social network such as Twitter is around 3 [73], which is similar to hateful followers' networks. It does not appear that hateful followers' networks are significantly more connected than the comparator risky network or Twitter networks on average. The overall degree of in-degree and out-degree distributions is illustrated in Figures 5.3, 5.4 and 5.5, showing long-tail characteristics where the majority of users follow a few numbers of the hateful accounts, from 1-10 (out-degree); and have very few followers, from 1-10 (in-degree).

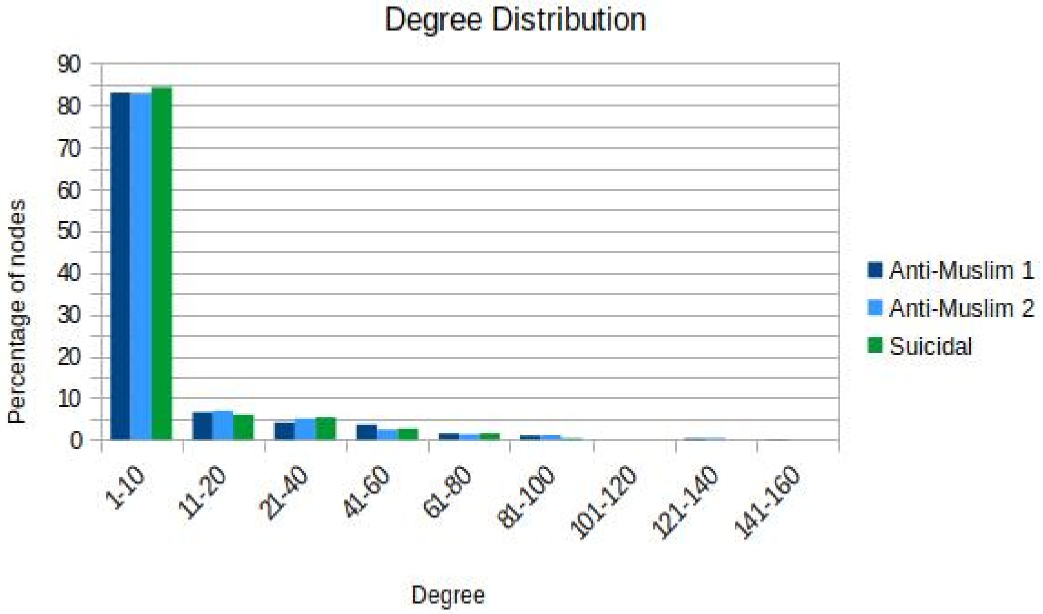


Figure 5.3: Degree distributions for the followers networks

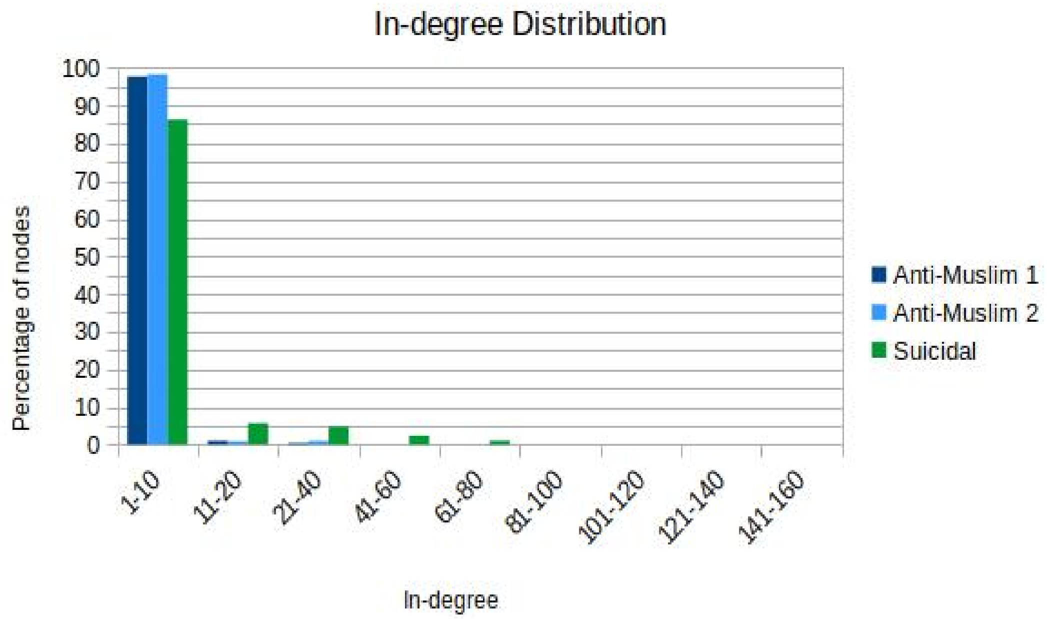


Figure 5.4: In-degree distributions for the followers networks

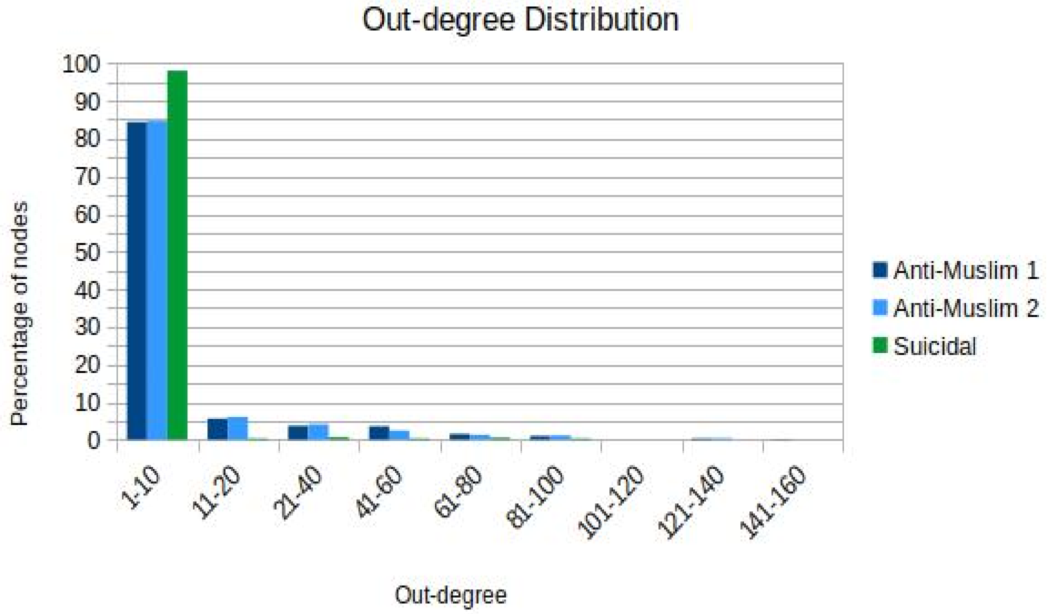


Figure 5.5: Out-degree distributions for the followers networks

This observation indicates the existence of hubs, i.e. a few nodes that are highly connected to other nodes, in hateful and suicidal networks. Hubs are nodes with a number of links that greatly exceed the average. These hubs have a power in terms of information spread and exposure. The presence of large hubs results in degree, in-degree and out-degree distributions with long tails. In a followers' network, hubs (highly connected nodes) are either an important influence or (can be influenced by) over network.

The distribution of the hateful followers' networks, in Figure 5.4, shows consistently similar in-degree distribution. Also, Figure 5.5 shows similar out-degree distribution for the hateful followers' networks. This could be further evidence of the structural similarity among the hateful followers' networks in these networks' samples.

Figure 5.4 shows a higher percentage of nodes in the far end of the in-degree tail for the suicidal network than hate networks. Conversely, Figure 5.5 shows more nodes in the far end of the out-degree tail for hateful networks, compared to the suicidal network. This suggests both hateful networks have fewer 'influencers' (people with lots of followers), and more 'super-consumers' (people who follow a lot of hateful posters). This could be interpreted as meaning that the hateful followers network tends to be more vulnerable to hate exposure [199] than the suicidal follower network, due to more nodes falling into the 'consumer' category and following a larger number of people creating risky content.

- **Average Clustering Coefficient:** The average clustering coefficients of the Anti-Muslim 1 and Anti-Muslim 2 followers' networks were 0.062 and 0.065, respectively. Clustering coefficient values for the hateful followers' networks were similar to the suicidal followers' network. Even though the average clustering coefficient of the hateful networks is similar to that of the suicidal network, their distribution showed that there are some differences. Empirically, nodes with higher degree(d_i) have a lower *local* clustering coefficient on average; thus, the *local* clustering coefficient (c_i) decreases with increasing degree [234]. The metric quantifies how close the neighbours are to being a complete graph (a clique). The distribution of clustering coefficients for the hateful followers' networks and the comparator network is shown in Figure 5.6.

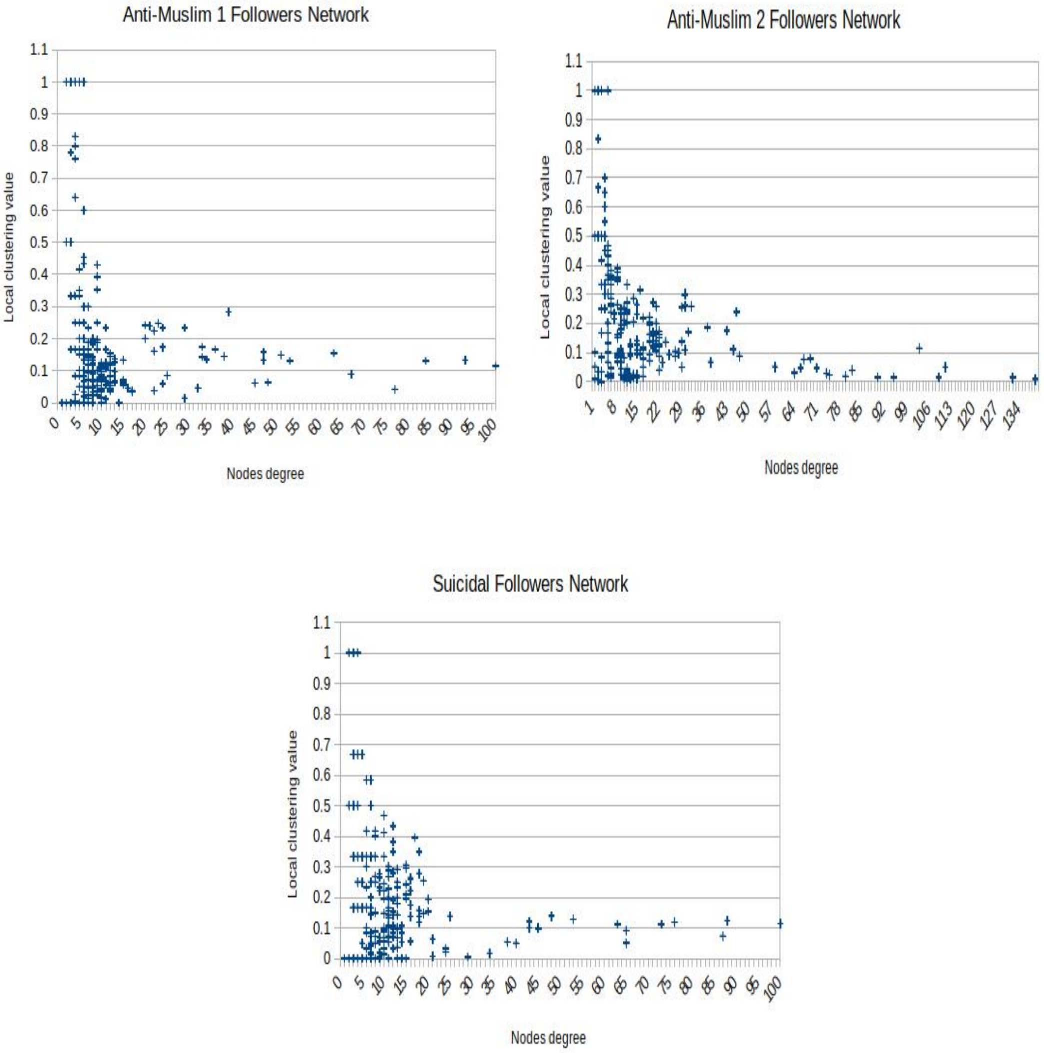


Figure 5.6: The distribution of the local clustering coefficient per degree for the followers' networks.

For the hateful followers' networks, several nodes with $(d_i) \geq 30$ have (c_i) greater than 0.2, while for the suicidal followers' network, all (c_i) of nodes that have $(d_i) \geq 30$ do not exceed 0.15. The probability of a node's neighbours also being connected (densely connected neighbours) is consistently higher for the hateful network than the suicidal network. Whether these nodes are 'hate-consumers' (out-degree edges) or 'hate-influencers' (in-degree edges), both cases exhibited densely connected neighbours. This behaviour was slightly lower in the suicidal followers' network, providing further evidence of tight connectivity between the hateful users, which means there is an increased risk of the hateful content exposure.

- **Reciprocity:**

  Table 5.3 shows that around 33% and 26% of the Anti-Muslim 1 and Anti-Muslim 2 followers' edges were reciprocal. These percentages are significantly lower than those recorded for the suicidal network, which was 62%. The presence of reciprocity in the followers' networks means that people with common interests, known as 'homophily', are exposed to each other's content [283, 270, 260]. Thus, about a third of the hateful accounts are exposed to each other's content, while more than 60% of the suicidal users do so. This suggests that the suicidal users form a more cohesive community based around reciprocal follower relationships [261, 268]. However, research has shown that reciprocity has a connection with 'emotional distress' which is significantly associated with suicidal users [232]. Thus, a study that investigates the incentive for reciprocal behaviour for different 'risky' followers' networks is required for clearly understanding that behaviour.

- **Diameter:**

  Table 5.3 shows that Anti-Muslim 1 and Anti-Muslim 2 recorded similar diameters (16 and 17) and average shortest paths (5.4 and 5.6). The maximum diameter is susceptible to outliers [253] and the average shortest path is a more rational measurement than the diameter because the diameter decreases when edges are added, while the latter may remain unchanged. Thus, the focus was on the average shortest path for characterising the hateful network. The average shortest paths are for the largest connected component (Giant Component) [196]. For example, in Anti-Muslim 1, between 5 and 6 steps are needed (5.4 avg. sh. path) to reach up to 60% of people who belong to the Anti-Muslim 1 followers' network (Giant Component). The bigger both metrics are, the easier the content will flow through the network. The metrics are similar for all three networks, though for suicide the Giant Component is smaller so fewer people are reached. Although the average path length should actually decrease with small sized networks [198], the average shortest path for the hateful followers' networks runs in line with that reported in a public retweets Konect dataset (5.45), which depicts a much more extensive Twitter network of online communications, with three million nodes and over ten million edges [186]. This provides some evidence that the hateful followers' networks exhibit data flow properties resembling large-scale communication followers' networks, albeit in a very small-scale network.

5.3.2 Retweets Graphs: measure of the hateful content contagion

- **Giant Component:** Table 5.4 shows that more than 69%, 81% and 72% of the nodes in the hateful retweets networks exist in the largest (giant) component. While some fluctuations exist between the hateful networks, the percentage in the comparator suicidal retweet network measures only 31.3%, which is significantly lower. These percentages reflect the percentage of users who are part of a connected community. The reason behind the large GC is the presence of the weak links which represent more tenuous acquaintance connections bringing together different groups of individuals. The 'weak link hypothesis', by Granovetter *et al.* [135], describes a specific social network structure in which strong links are associated with dense neighbourhoods (communities or groups), while weaker links act as bridges between them; the greater the number of weak ties, the greater the number of connected subgroups that make a bigger group. The results suggest that there is a consistently and significantly greater reach of content (contagion) in the hateful networks. Granovetter *et al.* [135] argues that contacts maintained through weak links play the important role of holding together groups, thus providing access to novel information. In that study, the seminal idea emerged of using node connectivity as an indicator of social network information spreading, an insight that is formalised in the GC notion [22]. In other words, GC connectivity, supporting the overall information spreading in social systems, would be most threatened by the removal of the node that is directly connected to the weak links [135]. Further research with much larger representative samples of networks is required to identify whether or not this pattern is generalisable beyond this study.

- **Density:**

  Table 5.4 shows that the densities of all the hateful retweets' networks in Table 5.4 are higher than 0.0005. Moreover, the densities of all the hateful retweet networks are higher (by 0.00033) than the suicidal network. Although Anti-Muslim 2 recorded the lowest density of all the hateful retweet networks, its density is higher than that of the suicidal retweet network. According to the sizes of the GC of the hateful networks, it is expected that the suicidal network will record a higher density than the hateful networks; this is because the density of the real systems network (e.g. Twitter networks) is inversely proportional to the size of that network or GC in this case [114]. However, Table 5.4 shows that the density of the suicidal's GC is also smaller than the ones recorded for the hateful networks. This means that the GC of the suicidal network has a smaller number of node connections than the hateful networks. Broadly speaking, the density measurements suggest higher connectivity among the hateful users compared to the suicidal retweets' network, which is interpreted as higher levels of information propagation. However, further research with larger representative samples of networks is needed to identify whether or not this pattern is generalisable beyond this study.

- **Average Degree:**

  Table 5.4 shows the average degree of the anti-Muslim networks (2.09, 2.3) is slightly higher than the antisemitic network (1.9), with all hateful networks higher than the comparison suicide network (1.4). The consistently higher average degree indicates more nodes were reachable on average and increased the propagation of hateful content through the network. The maximum degree of hateful content is higher than that of the suicide network for all three hateful retweet networks. The overall degree, in-degree and out-degree distributions are illustrated in Figures 5.7, 5.8 and 5.9 show a property of scale-free networks with the existence of fewer nodes in the network with higher levels of retweets, and many other nodes with fewer retweets.

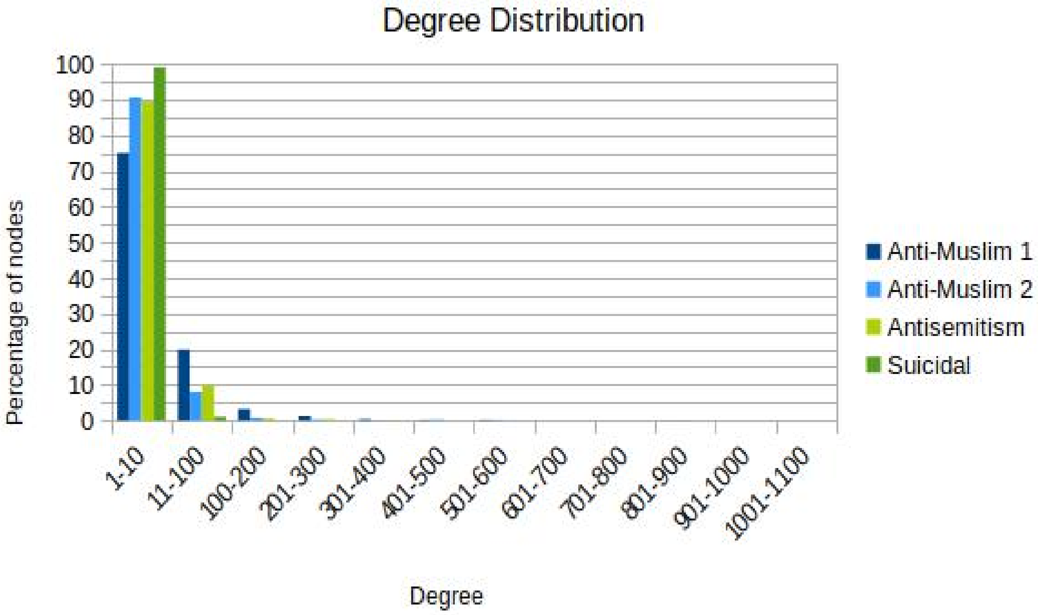


Figure 5.7: Degree distributions of the retweet networks

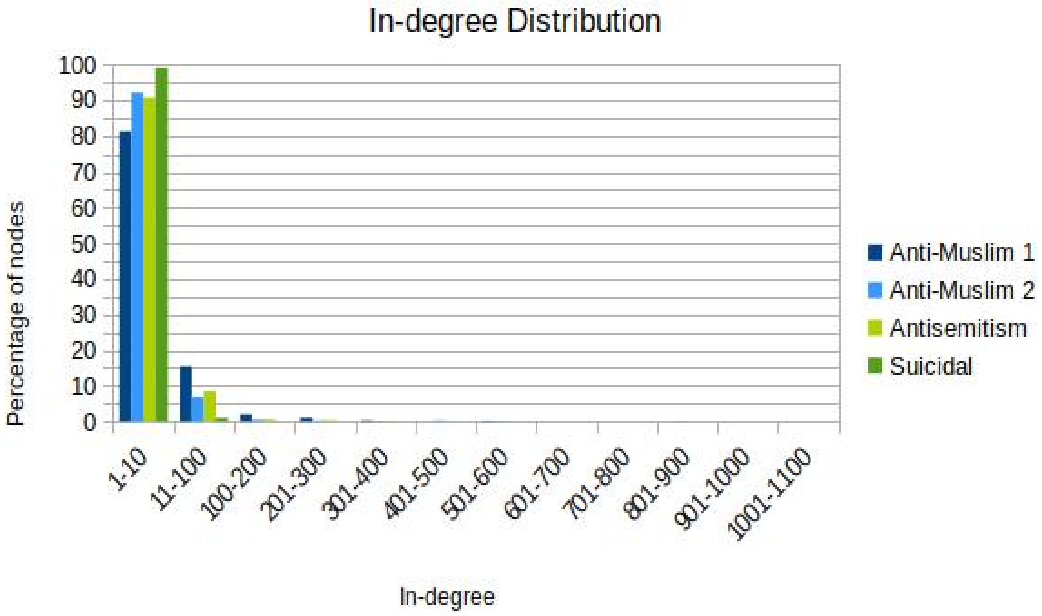


Figure 5.8: In-degree distributions of the retweet networks

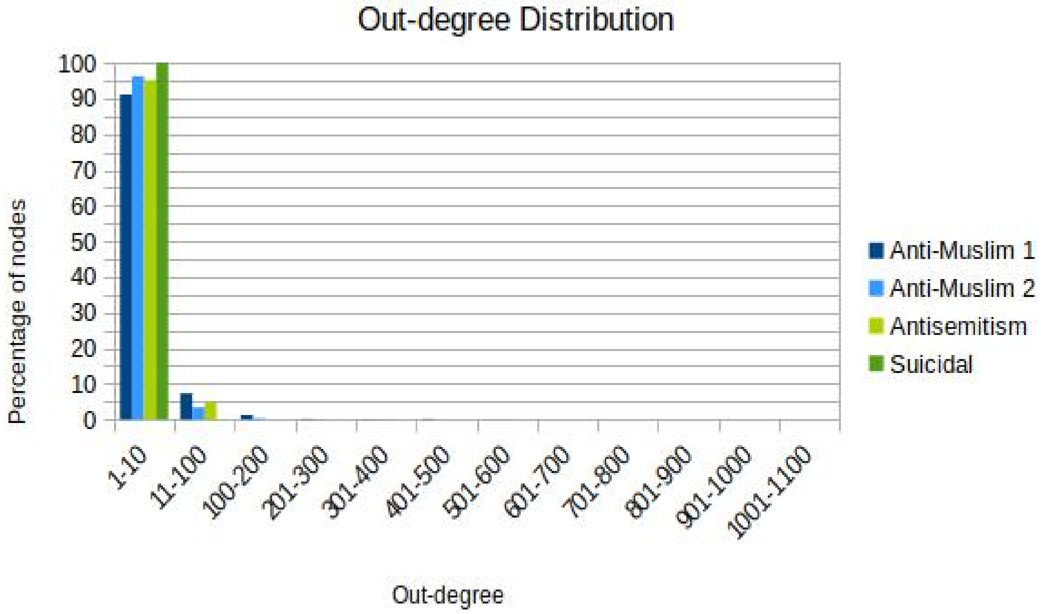


Figure 5.9: Out-degree distributions of the retweet networks

Moreover, all the hateful retweets' networks in this study show higher in-degree distribution than out-degree distribution for degrees larger than 10. This suggests that popular users (high in-degree), in this sample, are responsible for creating information cascades as they are highly retweeted by other hateful users in the hateful retweets' networks. This behaviour has also been seen in the suicidal retweets' network. Moreover, the suicidal retweets' network, in this sample, shows an absence of nodes with a high out-degree, compared to the number of out-degree edges in the hateful retweets' networks. Out-degree suggests a considerable number of the hateful users engaged significantly with hateful conversations by retweeting other users' hateful messages as 'super-retweeter', while the suicidal retweets' network does not indicate this behaviour. This suggests more co-operation in terms of the spread of the message (hate) in hateful networks, across all three hateful networks, and less in the suicide network would be seen.

Table 5.4 shows that the average clustering coefficient is low for all networks, suggesting a lack of coherent sub-groups within the overall networks. In addition, it is not feasible to show here the degree distribution of the *local* clustering coefficient as not all the networks exhibit a similar average clustering coefficient.

- **Reciprocity:**

  Table 5.4 shows that the reciprocity among the hate networks (12-18%) is higher than the comparator suicide network (0.9%) and fairly consistent. While the density and average clustering coefficients do not suggest significant smaller 'organised' subgroups exist (where users all retweet each other), there is clearly a consistent level of reciprocal retweeting in the context of hateful content that would appear to be higher than a comparator network. Although the relationship between these retweeters is not necessarily a friendship link, there is a noticeable level of co-operation.

- **Diameter:**

Table 5.4 shows there is a consistent level of information flow among the retweet networks, with the anti-Muslim network exhibiting a slightly higher diameter, but the antisemitic network having a slightly higher average shortest path. However, we should consider the size of the GC as the average shortest path is calculated for this largest sub-graph and not for the entire network. Table 5.4 shows that the hateful networks have a consistently sized (69%-81%) and significantly larger GC than the suicidal network (31.3%). This means while the same number of steps are needed to reach the largest cluster, the hateful content reach is consistently much larger than the suicide content - with between 37.9 and 50% more users reached by retweeting.

5.4 Conclusion

This chapter includes an analysis of the graph characteristics of three Twitter datasets of users who have posted tweets that human annotators agreed should be classified as containing evidence of hateful content. For the purposes of the research, these networks are referred to as hateful networks. Social network analysis (SNA) was conducted by investigating the social graphs of the followers and retweets of hateful users. Six metrics were applied on the hateful networks to examine the similarity and the differences between them, and results were compared with another "risky" network - suicidal ideation language.

For the hateful followers' networks, it was found that users were at similar levels in terms of risk of exposure to hateful content, with suicide networks as a comparator risky network at least 10% lower. This means that hateful followers' networks, in this sample, consistently tend to be more vulnerable to hate exposure [199] than the suicidal followers' network, which suggests a potential virality of hateful content.

For the hateful retweets networks, several structural similarities were observed in the sample, as were differences between the hateful retweet network in terms of social network metrics. Also, there was a consistently and significantly greater reach of content (contagion), and greater degree of co-operation on the spread of the message (hate) in hateful networks, across all three hateful networks, and less in the comparator 'risky' network - suicidal ideation. Hateful content reaches more users in fewer hops.

In the end, the aim of this study was to understand the characteristics of hateful online social networks in a non-representative small sample to help understand individuals' exposure to hate and cyberhate propagation. The findings provide some evidence that shows consistent metrics between hateful follower and retweet networks, many of which show higher risk of content exposure and contagion than a comparator 'risky' network of a similar size and connectivity. An expansion of this study using a much larger representative sample of networks could help the researcher and decision-makers understand and predict these networks' behaviour and suggest related solutions.

Naturally, our approach presents a limitation. It was conducted on a limited size set of annotated posts. Thus, it is still likely to represent a sub-section of the entire sample of hateful networks and hateful users engaged in posting hateful tweets, suggesting that our findings may be applicable to a small sample. However, because real-world networks such as social networks and the Internet are huge, obtaining the entire network structure is difficult, and typically only part of the network structure is available from network [132, 211, 218]. Also, even if the investigation started from a relatively large dataset, the users who post tweets classified as containing hate may face a suspension, which is a problem that has been faced in this study. When an account that posts hate speech is suspended, it is not possible to download the user's profile to collect their followers.

Research often refers to cyberhate as a virus that spreads like an infectious disease in our societies, affecting the most vulnerable people [120, 86]. One way to mitigate a specific virus's pandemic is to isolate the virus to prevent it from spreading. Similarly, cyberhate could possibly be managed by disrupting its propagation. One way to do this could be by removing/blocking users who post hateful ideologies to reduce the propagation of their content. This brings us to the question of which nodes could be most usefully removed to reduce the connectivity of a hateful network and decrease the propagation of hateful content. The next chapter considers different strategies for removing hateful users to reduce hateful networks' connectivity.

Chapter 6

Disrupting Hateful Networks

6.1 Introduction

The previous chapter focused on the characterisation of hateful networks to better understand their connectivity on Twitter. It showed that hateful followers' networks, in a sample at hand, exhibit connectivity characteristics consistent across them all and higher connectivity than the comparator risky network, suggesting a higher exposure to, and potential virality of, hateful content. Furthermore, it provided evidence that hateful retweeters in the sample display similar characteristics regarding some aspects, whilst they are different in other aspects. Additionally, they exhibit consistently high levels of information propagation behaviour, with a greater reach of content contagion in the hateful retweet networks than the comparator 'risky' network.

Over the past few years, the incidence of cyberhate reports has been increasing [2], suggesting an unsettling direction for the well-being and safety of our society. Furthermore, exposure to and engagement with online hate on social media has been suggested to promote offline aggression, with some perpetrators of violent hate crimes reported to have engaged with such content [102]. Online social media platforms have enabled social groups to reach much further into society than ever before. Online social networking sites have made it possible to spread different types of material, even transnationally, thereby providing additional avenues for encouraging hate and radicalism and allowing various hate groups to flourish online [245]. In addition, online hate groups and other individuals disseminating hateful ideologies can recruit young members to join or support their actions [193]; they may also directly incite extreme violence [51, 176]. Previous studies of youth exposure to online hate material show that visiting hate sites is associated with violent behaviour [348].

Online social media platforms are challenging to regulate [177], and policymakers have struggled to suggest practical ways of reducing cyberhate [7]. Efforts to ban and remove hate-related content have proved ineffective [1]. This is because removing certain content from a particular online source cannot guarantee the unavailability of the same content elsewhere [214]. Also, regulating hateful content may be considered to be against freedom of expression, e.g. cyber-libertarians argue against the regulation and censorship of internet content which could obstruct the free flow of ideas and information. Thus, we need to control or disrupt the connectivity and the flow of cyberhate and reduce exposure to and the propagation of hateful content through targeted intervention in the flow of hate. The aim of this chapter is to simulate disruption methods, also called node removal strategies [160, 53], prevention strategies [259] or intervention strategies [343], in order to curtail and contain cyberhate (through network pruning) on Twitter. Disruption methods aim to decrease or manage the connections or ties between the actors in a network. It is through these connections that actors can possess strategic positions, exchanging and sharing resources with other actors in the network. Disruption methods could include the possibility of identifying contagion pathways in hateful networks and evaluating the reduction in exposure of the network's users, also called the network's nodes (see Chapter 2, Section 2.6) to receiving hateful content, in the same way we might expect the spread of a traditional offline virus to be contained.

This study addresses the question of whether removing a proportion of nodes, depending on a specific role of that node in the hateful network, may affect network connectivity. Note that the followers' networks are conceptually different from the retweets networks, in that the former indicate exposure to hateful content, while the latter indicates the spread of content (contagion).

Examples of a specific role of a node in a hateful network are ’influencers’, ’super-consumers’ and ’super-retweeters’ as shown in the previous chapter. The previous chapter demonstrated that the hateful followers’ networks have fewer ’influencers’ and more ’super-consumers’ (people who follow a lot of hateful posters). In contrast, the hateful retweets’ networks have more ’influencers’ and fewer ’super-retweeters’ (people who retweet a lot of hateful posts). Accordingly, this assumes that removing the ’super-consumers’ may reduce the cyberhate exposure in the followers’ networks, while removing the ’influential’ users may decrease the content propagation of the hateful retweets’ networks - a practical application of this assumption would be useful to show the effect of these users on the network.

Thus, the assumptions, in general, are that (i) removing the crucial nodes (or high centrality nodes) from the followers’ networks would reduce the *exposure* to hateful content for other users; and (ii) removing the strategically positioned nodes (or high centrality nodes) from the retweets’ network would decrease the level of information propagation and *contagion*. These assumptions are applied to the generic (also called simple or one-mode) networks when the networks have one set of nodes that are similar to each other. Also, they are applied to the bipartite (also called affiliation or two-mode) version of the hateful networks which have two different sets of nodes and ties existing only between nodes belonging to different sets. Changes to the network connectivity (whether the structure of the network has significantly changed or ’disconnected’) can be measured using the common network connectivity metrics that were discussed in the previous chapter.

To the best of the author’s knowledge, it appears that no such study has applied disruption strategies to Twitter hateful networks to reduce network connectivity (exposure reduction) and potentially diffuse hate (contagion reduction).

The remainder of the chapter is organised as follows. Section 6.2 describes the methods of the present research. Section 6.3 reports the results, and the discussion and Section

6.5 concludes.

6.2 Methods

This section is reflecting the design and development/demonstration phases of the Data Science Research Methodology (DSRM) that have been mentioned in Chapter 3. Accordance to 3, this section aims to develop strategies (using the same datasets designed in Chapter 4) to identify nodes within hateful networks (user accounts) whose removal is empirically shown to reduce connectivity (largest component, density and average shortest path) in both the follower and retweet networks.

6.2.1 Data

To directly evaluate the effectiveness of disruptive strategies, the author used the same datasets used in the previous chapter. Two followers' networks: Anti-Muslim 1 and Anti-Muslim 2, and three retweets' networks: Anti-Muslim 1, Anti-Muslim 2, and an antisemitism network, in addition to the comparator risky networks: suicidal followers' networks and suicidal retweets' networks (see Chapter 5, Section 5.2.1). To remind the reader, Table 6.1 summarises the number of nodes and edges for each network.

Table 6.1: Numbers of the nodes and the edges of the hateful networks

Followers' networks		
Network	**Nodes**	**Edges**
Anti-Muslim 1	1004	2644
Anti-Muslim 2	1073	2895
Suicidal	987	2410
Retweets' networks		
Network	**Nodes**	**Edges**
Anti-Muslim 1	1229	2571
Anti-Muslim 2	5581	16338
Anti-Semitic	2748	5091
Suicidal	3209	2211

6.2.2 Centrality Metrics

To understand the importance of the network's nodes being studied, it is normal to start with an evaluation of their location relative to all other nodes in the network [206]. In graph theory and network analysis, indicators of centrality identify the most important nodes within a graph. One node (or a set of nodes) is considered to be important if its removal would largely reduce the network's stability or robustness or make the network more vulnerable [53, 251].

There are many measures of node centrality that capture the importance of nodes within a network. This research is interested in five centrality metrics: the degree centrality and the separate measures of degree centrality, namely in-degree and out-degree; the betweenness centrality, and the eigenvalue centrality of a given node [324, 70] or a group of nodes. The measures are selected because they are frequently used centrality measures for removal strategies [112] (see also Table 2.7, Chapter 2).

- **Degree centrality:** The degree centrality for a node is simply its degree. Recall that a node's degree is simply a count of how many social connections (i.e., edges) it has. A node with 10 social connections (edges) would have a degree centrality of 10. A node with 1 edge would have a degree centrality of 1. For a graph G with n nodes and edges e, the degree centrality for a node $v \in G$ is $d(v)$:

$$d(v) = \frac{degree(v)}{n-1} \tag{6.1}$$

The degree can be interpreted in terms of the immediate risk of a node catching whatever is flowing through the network (such as a virus, or some information). Thus, nodes with high degree are considered important in network exposure and propagation. In a theoretical study, Golub *et al.* [134] found that the efficient diffusion of influence through a network is limited by the presence of highly influential, high degree nodes. In the case of a directed network (where ties have direction, e.g. followers' and retweets' networks), we usually define two separate measures of degree centrality, namely in-degree and out-degree [292]. Accordingly, in-degree is a count of the number of ties directed to the node, whereas out-degree is the number of ties that the node directs to others.

In terms of follower networks, out-degree means that users can directly consume (see, read) the content posted by other users. In this case, the research interest lies in the out-degree, which represents the number of users that someone follows (e.g. if A has an out-degree of 5, it means A follows five people). Higher out-degree values mean a wider exposure to different sources of information propagators (more information coming from a number of other sources). Also, the research interest lies in the in-degree, which represents the number of followers that follow a user (e.g. if A has an in-degree of 5, it means five people follow A). Higher in-degree value refers to influential users (content creation hubs or conversational hubs) who can be responsible for hate creation and propagation.

For the retweet network, in-degree for each node is how many other users retweeted the tweet of the user (e.g. if A has an in-degree of 5, it means they have been retweeted by five different users). Out-degree represents the number of retweets a user tweet from other users (e.g. if A has an out-degree of 5, it means they retweeted five tweets posted by five different users). High in-degree indicates high hate propagation as more users are retweeting the specific member's tweets; hence, this member has more influence on others [162]. On the other hand, high out-degree indicates a high level of diversity of the retweeted content as a user is retweeting a large number of accounts. They may only consume information and not be creating it or influencing others in a significant way [21]. Their importance is also because they can interact with many others [142]- e.g. in collaboration networks, this measure helps indicate how collaborative each author is [340]. According to Roland *et al.* [275], in retweet networks, in- and out-degree centrality metrics capture the users' engagement with other users and the content of their posts; and they also form vital bridges. These metrics indicate the actual attention given to content and the action that users took to disseminate information. So, both in-degree and out-degree nodes could play an essential role in the networks that are examined here.

- **Betweenness centrality:** This metric is a measure of accessibility that is the number of times a node is crossed by shortest paths in the graph. It represents the degree to which nodes stand between each other. In other words, the more people depend on a user to make connections with other people, the higher that user's betweenness centrality becomes. For example, in a telecommunications network[1], a node with higher betweenness centrality would have more control over the network, because more information will pass through that node. This is useful for finding the individuals who influence the flow around a system. For a graph G, the betweenness centrality for a node $v \in G$ is $b(v)$:

$$b(v) = \sum_{s \neq t \neq v)} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{6.2}$$

where σ_{st} denotes the number of shortest paths between nodes s and t (usually

[1] A telecommunications network is a group of nodes interconnected by links that are used to exchange messages between the nodes, e.g. computer networks and the Internet

σ_{st} = 1) and σ_{st} (v) expresses the number of shortest paths passing through node v .

- **Eigenvalue centrality:** A node with a high eigenvalue is important as a connector for high information diffusion. Degree centrality measures the number of connections a node has but disregards the position of the nodes to which they are connected. Eigenvalue centrality modifies this approach by giving a higher centrality score to nodes with connections to other nodes that are themselves central. In other words, it can be said that the centrality of a given node *v* is proportional to the sum of the centralities of *v* 's neighbours. This is the assumption behind the eigenvalue centrality formula, which is for a graph G and let let $A = (a_{v,j})$ be the adjacency matrix[2], the eigenvalue centrality for a node $v \in G$ is $e(v)$:

  $$e(v) = x_{v\frac{1}{\lambda}} \sum_{j=1}^{n} a_{vj} x_j \tag{6.3}$$

  where x_v , x_j denotes the centrality of node *v, j* and n denotes the number of the graph nodes. λ denotes the largest eigenvalue of G and a_{vj} represents an entry of the adjacency matrix $A(a_v, a_j)$ = 1 if nodes *v* and *j* are connected by an edge and $(a_v, a_j$) = 0 otherwise.

The closeness and betweenness centralities embody closely related ideas [206]. The closeness centralities measure the inverse of the sum of all shortest paths to other nodes. However, the closeness centrality metric was excluded because it measures the closeness of a node in the biggest component rather than in an entire network. If the biggest component is highly connected, all the nodes would be shown with a similar score, meaning that all the nodes would have a high closeness.

[2] In graph theory, an adjacency matrix is a square matrix used to represent a finite graph. The elements of the matrix indicate whether pairs of vertices are adjacent or not in the graph.

6.2.3 Node Removal Strategies

The previous section explored several centrality methods used to detect the importance of a node in a network. The challenge here is to identify nodes within the hateful networks, whose removal would decrease the connectivity and reduce the flow of hateful content. The followers' networks indicate exposure to hateful content, while the retweets' networks indicate the spread of content (contagion). Therefore, the assumptions are that: (i) removing the important nodes (high central nodes) from the followers' networks would reduce the *exposure* to hateful content for other users (remove key content providers); and (ii) removing the important nodes (high central nodes) from the retweets' network would decrease the level of information propagation and *contagion*. To efficiently identify nodes v whose removal reduces exposure and contagion within the network most, 13 structured heuristic node removal strategies were designed using different node centrality metrics explained below:

Single Node Removal Strategies

- **Degree-based strategy:** This strategy uses a degree centrality metric to select the nodes with the highest degree max_{deg}(v) and remove them. The steps of the degree-based strategy are simply explained in Algorithm 6.1.

Algorithm 6.1 Degree-based Node Removal

```
INPUT: $H$=$v_1, v_2, .., v_h$
OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H
repeat until remove ($10\%$ OF$H$)
  for each $v_i \in$ H do calculate degree centrality
    if v $= max_{deg}$ then
      Remove $v$
      if ($1\%$ OF$H$) nodes are removed then
        Return $GC(H), d(H), l(H)$
        Recalculate degree centrality
      else
        Remove $v$
    end if
  else
    Return $null$
  end if
  end for
```

- **Indegree-based and Outdegree-based strategies:**

  These strategies select the nodes with the highest in-degree max_{indeg}(v) or highest out-degree max_{outdeg}(v) and remove them. The equation used to calculate the in-degree and out-degree centrality is the same as the equation for degree centrality, but it considers out-degree or in-degree instead of degree.

  Algorithm 6.2 explains the in-degree-based strategy and; Algorithm 6.3 explains the out-degree-based strategy.

Algorithm 6.2 Indegree-based Node Removal

INPUT: H=$v_1, v_2, .., v_h$

OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H

```
repeat until remove (10$\%$ OF$H$)
  for each $v_i \in$ H do calculate indegree centrality
    if v $= max_{indeg}$ then
      Remove $v$
      if (1$\%$ OF$H$) nodes are removed then
        Return $GC(H), d(H), l(H)$
        Recalculate indegree centrality
      else
        Remove $v$
    end if
  else
    Return $null$
  end if
  end for
```

Algorithm 6.3 Outdegree-based Node Removal

INPUT: H=$v_1, v_2, .., v_h$

OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H

```
repeat until remove (10% OF$H$)
  for each $v_i \in$ H do calculate outdegree centrality
    if v $= max_{outdeg}$ then
      Remove $v$
      if (1% OF$H$) nodes are removed then
        Return $GC(H), d(H), l(H)$
        Recalculate outdegree centrality
      else
        Remove $v$
    end if
  else
    Return $null$
  end if
  end for
```

- **Betweenness-based strategy:** This strategy selects the nodes with the highest betweenness max_{bet}(v) and removes them - see Algorithm 6.4.

Algorithm 6.4 Betweenness-based Node Removal

INPUT: H=$v_1, v_2, .., v_h$

OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H

```
repeat until remove (10% OF$H$)
    for each $v_i \in$ H do calculate betweenness centrality
        if v $= max_{btw}$ then
            Remove $v$
            if (1% OF$H$) nodes are removed then
                Return $GC(H), d(H), l(H)$
                Recalculate betweenness centrality
            else
                Remove $v$
        end if
    else
        Return $null$
    end if
    end for
```

- **Eigenvalue-based strategy:**

  This strategy selects the nodes with the highest eigenvalue max_{eig}(v) and removes them -see Algorithm 6.5.

Algorithm 6.5 Eigenvalue-based Node Removal

```
INPUT: $H$=$v_1, v_2, .., v_h$
OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H
repeat until remove ($10\%$ OF$H$)
  for each $v_i \in$ H do calculate eigenvalue centrality
    if v $= max_{egv}$ then
      Remove $v$
      if ($1\%$ OF$H$) nodes are removed then
        Return $GC(H), d(H), l(H)$
        Recalculate eigenvalue centrality
      else
        Remove $v$
    end if
  else
    Return $null$
  end if
  end for
```

- **Random-based strategy:** Random node removal strategy was used as a baseline to examine the performance of the five structured node removal strategies. This strategy selects the nodes randomly and removes them. It is also important to investigate if the hateful networks are robust against random node removal or if they show different behaviour to scale-free networks [23, 22, 209, 91].

The fundamental differences between the degree-based, the betweenness-based and the eigenvalue-based strategies are that the degree-based method concentrates on reducing the total number of edges in the network as fast as possible, whereas the betweenness-based approach concentrates on removing as many edges as possible in the shortest path [150]. Holme *et al.* [150] showed that the degree-based procedure removes edges connecting nodes with high degrees very fast; as a consequence, the original network maybe split into many subgraphs of nodes with low (but not zero) degrees. The betweenness-based strategy, on the other hand, concentrates on nodes of high betweenness, and thus the nodes that passed by more shortest paths are first lost. Consequently, it may elongate the average shortest path but not necessarily disconnect a network; thus, it indicates the potential of a node in controlling communications in a network. The eigenvalue-based strategy aims to deconstruct the bridges between the highest impact nodes. A node with a large eigenvector centrality is important because it is connected to nodes with more connections. Eigenvalue centrality was also used to measure the popularity and importance of a node in (non-hateful) social networks by Newman *et al.* [240] and Bonacich *et al.* [55].

Implementation was undertaken using the Python language, NetworkX package. As an example of a degree-based strategy, the degrees of the graph table were calculated and sorted in descending order. The first 1% of the node was removed then the GC was recalculated. The following code was used:

```
nx.degree(G)
sort= sorted(G.degree, key=lambda x: x[1], reverse=True)
L = nx.line_graph(G)
count=(1/L)
remove = [node for node,degree in
dict(G.degree()).items() if sort > count]
G.remove_nodes_from(remove)
largest_cc = max(nx.connected_components(G))
```

Hybrid Node Removal

The hybrid node removal strategy in this study was a combination of two strategies applied to the network as one strategy. The objective of applying these strategies was to provide further insight into node removal strategies in cyberhate networks. Seven hybrid strategies were examined: DegreeBetweenness-based, DegreeEigenvalue-based, IndegreeBetweenness-based, IndegreeEigenvalue-based, OutdegreeBetweenness-based, OutdegreeEigenvalue-based and BetweennessEigenvalue-based strategies. To avoid any bias that may occur in the combined (hybrid) node removal strategy's results, we decided to delete equal numbers of the highest centrality nodes of each strategy of the combined strategy. For more clarification, the node deletion for a DegreeBetweenness-based strategy is given below:

- (a):Calculate the degree and betweenness centrality of the network's nodes.
- (b):Search for the maximum degree nodes, arrange them in descending order and then delete 0.5% of H nodes.
- (c): Search for the maximum betweenness nodes, arrange them in descending order and then delete 0.5% of H nodes.
- (d): Now the overall deleted nodes is 1% of the H nodes.
- (e): Calculate the (GC), density and the average shortest path.
- DO this until 10% of nodes are deleted.

Seven hybrid node removal strategies are defined below:

- **DegreeBetweenness-based strategy:** This strategy selects nodes with the highest degree and highest betweenness and removes them. Algorithm 6.6

Algorithm 6.6 DegreeBetweenness-based Node Removal

```
INPUT: $H$=$v_1, v_2, .., v_h$
OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H
repeat until remove (10% OF$H$)
  for each $v_i \in$ H do calculate degree and betweenness centrality
    if v $= max_{deg}$ then
      Remove $v$ DO until (0.5% OF $H$) nodes are removed
      if v $= max_{btw}$ then
        Remove $v$
        if (1% OF $H$) nodes are removed then
          Return $GC(H), d(H), l(H)$
          Recalculate degree centrality and betweenness centrality
        else
          Remove $v$
      end if
    else
      Return $null$
    end if
    end for
```

- **IndegreeBetweenness-based strategy:** This strategy selects nodes with the highest indegree and highest betweenness and removes them. Algorithm 6.7 explains IndegreeBetweenness-based strategy as follow:

Algorithm 6.7 IndegreeBetweenness-based Node Removal

INPUT: H=$v_1, v_2, .., v_h$

OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H

```
repeat until remove (10% OF$H$)
 for each $v_i \in$ H do calculate degree and betweenness centrality
    if v $= max_{indeg}$ then
       Remove $v$ DO until (0.5% OF $H$) nodes are removed
       if v $= max_{btw}$ then
          Remove $v$
          if  (1% OF $H$) nodes are removed then
             Return $GC(H), d(H), l(H)$
             Recalculate indegree centrality and betweenness centrality
          else
             Remove $v$
       end if
    else
       Return $null$
    end if
    end for
```

-

- **OutdegreeBetweenness-based strategy:** This strategy selects nodes with the highest outdegree and highest betweenness and removes them. Algorithm 6.8 explains OutdegreeBetweenness-based strategy as follow:

Algorithm 6.8 OutdegreeBetweenness-based Node Removal

INPUT: H=$v_1, v_2, .., v_h$

OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H

```
repeat until remove (10% OF$H$)
 for each $v_i \in$ H do calculate outdegree and betweenness centrality
   if v $= max_{outdeg}$ then
     Remove $v$ DO until (0.5% OF $H$) nodes are removed
     if v $= max_{btw}$ then
       Remove $v$
       if (1% OF $H$) nodes are removed then
         Return $GC(H), d(H), l(H)$
         Recalculate outdegree centrality and betweenness centrality
       else
         Remove $v$
     end if
   else
     Return $null$
   end if
   end for
```

- **BetweennessEigenvalue-based strategy:** This strategy selects nodes with the highest betweenness and highest eigenvalue and removes them. Algorithm 6.9 explains BetweennessEigenvalue-based strategy as follow:

Algorithm 6.9 BetweennessEigenvalue-based Node Removal

```
INPUT: $H$=$v_1, v_2, .., v_h$
OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H
repeat until remove (10% OF$H$)
 for each $v_i \in$ H do calculate betweenness and eigenvalue centrality
    if v $= max_{btw}$ then
        Remove $v$ DO until (0.5% OF $H$) nodes are removed
        if v $= max_{egv}$ then
            Remove $v$
            if (1% OF $H$) nodes are removed then
                Return $GC(H), d(H), l(H)$
                Recalculate betweenness centrality and eigenvalue centrality
            else
                Remove $v$
        end if
    else
        Return $null$
    end if
    end for
```

- **DegreeEigenvalue-based strategy:** This strategy selects nodes with the highest degree and highest eigenvalue and removes them. Algorithm 6.10 explains DegreeEigenvalue-based strategy as follow:

Algorithm 6.10 DegreeEigenvalue-based Node Removal

```
INPUT: $H$=$v_1, v_2, .., v_h$
OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H
repeat until remove (10% OF$H$)
  for each $v_i \in$ H do calculate degree and eigenvalue centrality
    if v $= max_{deg}$ then
      Remove $v$ DO until (0.5% OF $H$) nodes are removed
      if v $= max_{egv}$ then
        Remove $v$
        if (1% OF $H$) nodes are removed then
          Return $GC(H), d(H), l(H)$
          Recalculate degree centrality and eigenvalue centrality
        else
          Remove $v$
      end if
    else
      Return $null$
    end if
    end for
```

- **IndegreeEigenvalue-based strategy:** This strategy selects nodes with the highest indegree and highest eigenvalue nodes and removes them. Algorithm 6.11 explains IndegreeEigenvalue-based strategy as follow:

Algorithm 6.11 IndegreeEigenvalue-based Node Removal

INPUT: H=$v_1, v_2, .., v_h$

OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H

```
repeat until remove (10% OF$H$)
 for each $v_i \in$ H do calculate indegree and eigenvalue centrality
   if v $= max_{indeg}$ then
     Remove $v$ DO until (0.5% OF $H$) nodes are removed
     if v $= max_{egv}$ then
       Remove $v$
       if (1% OF $H$) nodes are removed then
         Return $GC(H), d(H), l(H)$
         Recalculate indegree centrality and eigenvalue centrality
       else
         Remove $v$
     end if
   else
     Return $null$
   end if
   end for
```

- **OutdegreeEigenvalue-based strategy:** This strategy selects nodes with the highest outdegree and highest eigenvalue and removes them. Algorithm 6.12 explains OutdegreeEigenvalue-based strategy as follow:

Algorithm 6.12 OutdegreeEigenvalue-based Node Removal

```
INPUT: $H$=$v_1, v_2, .., v_h$
OUTPUT: Giant component (GC), density (d) and average shortest path (l) of H
repeat until remove ($10\%$ OF$H$)
 for each $v_i \in$ H do calculate outdegree and eigenvalue centrality
   if v $= max_{outdeg}$ then
     Remove $v$ DO until ($0.5\%$ OF $H$) nodes are removed
     if v $= max_{egv}$ then
       Remove $v$
       if ($1\%$ OF $H$) nodes are removed then
         Return $GC(H), d(H), l(H)$
         Recalculate outdegree centrality and eigenvalue centrality
       else
         Remove $v$
     end if
   else
     Return $null$
   end if
   end for
```

Networkx package of the Python language was used for implementation. The code is similar to the code applied to the single node-removal strategy; however, it calculates two centralities (e.g. degree and betweenness) for each node. After that deletion is implemented on 0.5% of the highest degree nodes and then on 0.5% of the highest betweenness nodes. Ultimately, the percentage of deleted nodes is 1% of *H* for the first stage of deletion. Then we calculate the networks' performance.

Only the first 10% of the highest central nodes were removed, and the results of the metrics were recorded gradually for each 1% removed. Previous research results show that highly influential nodes are rare in social networks [357], which is the reason we chose to only remove 10% in descending order of the centrality of the nodes. For instance, Otsuka *et al.* [248], Gallos *et al.* [124], Xu *et al.* [343] and Duijn *et al.*[108] considered removing 4%-8%-10%. The results of the removal strategies approximate the effects of different strategies that reflect the role of the node within the network.

In each round of node removal, the centrality metrics had to be recalculated because, according to Nie *et al.* [242], Bellingeri *et al.* [40], Cohen *et al.* [84]and Iyer *et al.* [158], this provides more efficient deletion than the non-recalculated method. Node removal strategies were applied to the hateful networks (followers' and retweets' networks), and also applied to the suicidal networks to show the similarities and differences in the role of the nodes within non-hateful networks.

Network performance metrics

It is expected that the hateful network will be restructured after removing a portion of specific nodes. The impact of different node removal strategies was measured through changes in the networks' GC, the density, and the average shortest path metric of the different networks, see Chapter 5, Section 5.2.3. We selected these metrics because network GC and network average shortest path have been widely used as an indicator of network changes/failure/distribution [343, 264, 52, 168] - also see Table 2.7, Chapter 2. GC represents the maximum number of nodes connected among the network and is the simplest and most widely applied indicator of network functioning - adopted to evaluate the connected-ness of internet routers [23], the vulnerability of power grids [345] or as a measure of the epidemic spreading, in terms of finding the best vaccination strategies [80, 282]. Reducing the GC to a small, connected component is a positive sign of the effectiveness of the node removal strategy.

In contrast, the increase of the average shortest path is a positive sign of the node removal strategy as it indicates the removal of vital bridges (fundamental hubs) that 'shorten' the distance between the nodes. Figure 6.1 shows a simple network containing numbered nodes as an example of a network being restructured after removing a node. The shortest path between $node_0$ and $node_2$ is 2, passing through $node_3$. When $node_3$ is removed, the shortest path between $node_0$ and $node_2$ will become 4, as the content needs to be passed via 3 nodes - $node_4$ then $node_5$, and then $node_1$ to reach its destination.

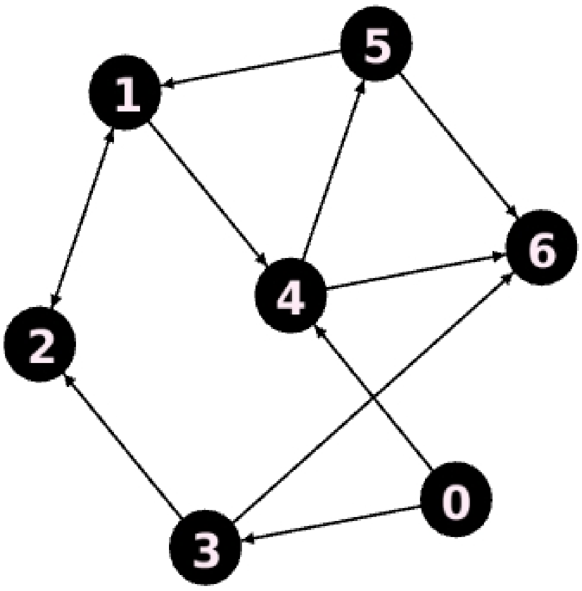


Figure 6.1: Example graph that shows removing $node_3$ will increase the shortest path between $node_0$ and $node_2$.

Also used is the density metric; this is because a previous study by Luarn *et al.* [207] showed that network density is positively related to transmitter activity on social network sites. Moreover, on Twitter, the rate at which information is spread through a network was found to depend on its density [197]. The clustering coefficient metric has been excluded because not all the hateful retweets' networks exhibit a similar average clustering coefficient. Also, the reciprocity metric has been excluded as it is very tiny for the suicidal retweets' network compared to the hateful retweets' network.

6.2.4 Removal Strategies on the Bipartite Networks

Most social networks are conceived of as relationships among a set of nodes; an example of this is our generic hateful followers and retweet networks previewed in Section 6.2.1 (also called one-mode as they are represented as a 1-mode matrix). However, bipartite network (also referred to as two-mode) data are also common in social network contexts. Bellingeri *et. al.*[40] found that the efficiency of removal strategies (fraction of nodes to be deleted for a given reduction of GC size) depends on the topology of the network. Applying bipartite networks, we can obtain different topology that may show different nodes' roles in a network- we need to detect different nodes' roles for finding the best way for disrupting the hateful networks. In generic networks (one-mode), user one is connected with user two by a tie that represents the relationship, (follow/followed) or (retweet/retweeted). In a bipartite network, user one is connected to user two by an affiliation relationship, e.g. user A and user B are both following the same user.

To make this clearer, we refer the reader to Figure 6.2 below:

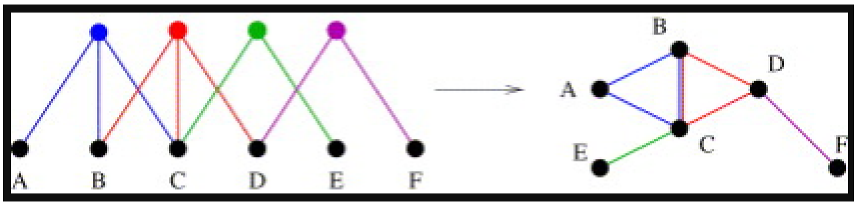


Figure 6.2: A bipartite network (left) and the generic network (right). Notice that the link is obtained twice since B and C have two neighbours in common in the bipartite network. .

Figure 6.2 shows a bipartite network (left) and the same network in a generic mode (right). It demonstrates that each top vertex induces a clique (complete subgraph) between the bottom nodes to which it is linked. Consider that B and C users follow the blue, but B and C may not know each other; they belong to the same affiliation *Blue* (they belong to the same clique). Tables 6.2 show the adjacency matrix of the generic network and the bipartite network.

Table 6.2: One-mode matrix for the generic network (left) and Two-mode matrix for the Bipartite network (right).

	A	**B**	**C**	**D**	**E**
A	2	1	1	0	0
B	1	3	1	1	0
C	1	1	3	1	1
D	0	1	1	3	0
E	0	0	1	0	0

	Blue	**Red**	**Green**	**Purple**
A	1	0	0	0
B	1	1	0	0
C	1	1	1	0
D	0	1	0	1
E	0	0	1	0

The two tables are an examples of the simple (generic) network in Table 6.2(left) and the bipartite version of these networks (right). Table 6.2 (left) shows that the:

- The rows and column are the same set.
- The numbers refer to how many groups (cliques) the one user shares with others, in terms of the follower's network, **A** share with **B** a follow/retweet relationship.
- The numbers mean that some ties are stronger than others. For example, **A** and **E** are not sharing the following relationship meaning that they may or may not be connected via a weak tie. The same goes for the retweet networks.

For the bipartite network, Table 6.2 (right) shows that the:

- The rows and the column are not the same set.
- The 0,1 numbers refer to whether the users in the column are affiliated with (or belong) to the group (clique) in the rows. In terms of the followers' network, **A** belongs to **Blue** group (clique) but not to **Red** group.
- The third row indicates that **C** belongs to the three relationship groups. The same goes for the retweet networks.

The two-mode or the bipartite network adds an extra layer for the formation of relationships among the network's users. This means that the important role of the users in the network may become slightly different in a bipartite network. For example, when applying an intended attack or (node removal) to a generic network, Figure 6.2 (right), we are targeting the nodes themselves, while for an intended attack on the bipartite network, Figure 6.2 (left), we consider the removal of the nodes in the bottom set and/or the group's node in the above set. However, there are differences between the values obtained from the bipartite models and real-world networks. In bipartite networks, many top nodes have a large neighbourhood intersection. In other words, the overlap between cliques is significant, and more precisely, if two cliques have one node in common, then they certainly have many[191]. So does this means that removing a common node may affect the overlapping cliques or will they still be connected by the other common nodes? This is the motivation behind applying the chosen node removal strategies to bipartite hateful networks. Bipartite networks may be extracted from a directed network or indirected network. Above we explained the bipartite network in general. In this study, we are interested in the directed bipartite, which is extracted from a directed network; thus, we first need to convert the directed networks into a directed bipartite version, similar to [358] and [41]. This is shown in Figure 6.3.

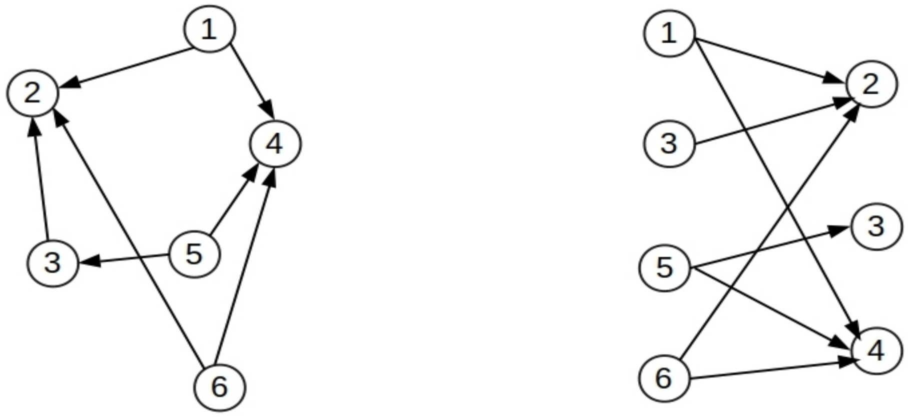


Figure 6.3: Constructing a bipartite graph from a directed one. Left: directed graph. Right: bipartite graph.

In Figure 6.3 (right) users of numbers 2, 3 and 4 are followed or retweeted by users of numbers 1, 3, 5 and 6. Users of number 1 and 3 follow/retweet the same users, meaning that they belong to the same affiliation relationship. Notice that user number 3 exists in both sets; however, it is considered as two users. When number 3 is removed from one set, it will not be removed from the other set.

Figure 6.4 shows the steps needed to apply node removal strategies to a bipartite network.

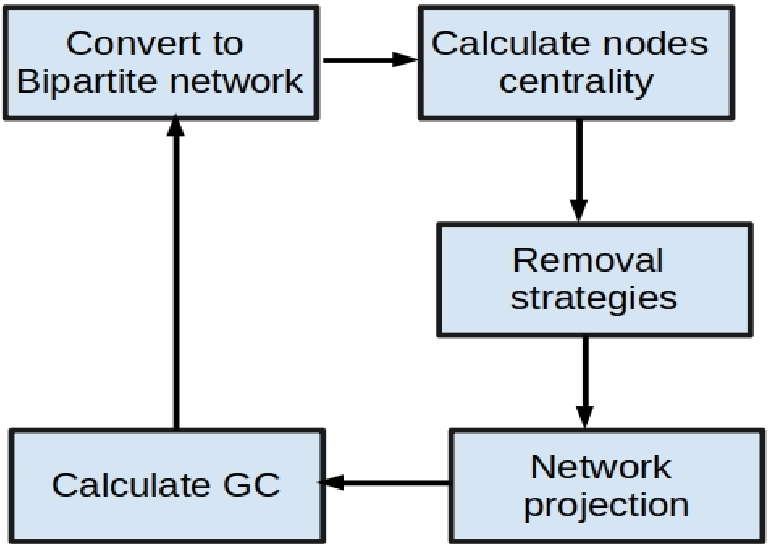


Figure 6.4: Steps of the node removal strategies on the bipartite networks

Figure 6.4 illustrates that we first converted the edgelist of our network using Panda package[3] Data frame into two-mode matrix , as follows:

```
df = pd.get_dummies(OurEdgelist.set_index('Target')
['Source']).max(level=0)
```

[3] Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations to allow the user to manipulate numerical tables and time series.

Then build the bipartite network using the Networkx package.

```
from networkx.algorithms import bipartite
B = nx.Graph()
# Add nodes with the node attribute "bipartite"
B.add_nodes_from(df.Target, bipartite=0)
B.add_nodes_from(df.Source, bipartite=1)
# Add edges only between nodes of opposite node sets
G = B.add_edges_from(df)
```

After that we calculated the node centrality for all the nodes on *B*. Then the node removal strategies were applied: 'single', hybrid and random node removal strategies; the codes are similar to those mentioned previously for the single and hybrid node-removal strategies. In order to compare the performance of the node removal strategies on the bipartite (two-mode) networks with the performance of the same strategies on the generic networks (one-mode), we converted the bipartite network after the removal of the nodes into a one-mode network using **projection**. Directed networks are allowed as input for the projection. The output will also then be a directed network with edges if there is a directed path between the nodes in the bipartite networks. The projection has been implemented using the NetworkX package as follows:

```
Import Networkx as nx
from networkX.algorithms import bipartite
G=nx.projected_graph(B,Target)
```

The reason for converting the bipartite network into a one-mode network in this step is because bipartite networks (two-mode networks) tend to have more and larger GCs than generic networks (one-mode networks)[324]. These are produced when three or more nodes are connected to a common node in the two-mode network. The performance of the node removal strategies was measured on only the largest component (GC), because its value is near to the largest component of the generic hateful networks and projected hateful networks. The density and the average shortest path were excluded due to their measurement differences between the generic and the projected networks (both one-mode networks). For example, the average shortest path for the Anti-Muslim 1 follower network was 5.4, while it was 5.9 for the projected network.

6.3 Results and Discussion

This part reflects the Data Science Research Methodology (DSRM) evaluation stage that was described in Chapter3. According to Chapter 3, this section aims to identify if applying node removal strategies (disruption strategies), depending on the node role in the network, will reduce network connectivity (exposure reduction) and diffuse the spread of hate (contagion reduction).

6.3.1 Single Node Removal Strategies

Followers Networks

Figures 6.5, 6.6 and 6.7 show the six removal strategies that were applied to the followers' networks. Of the six removal strategies tested, the random removal strategies were the least effective. This is in line with Jahanpour *et al.* [160], Crucitti *et al.* [90] and Wang *et al.* [319], who found that small-world networks have strong resilience against a random-based node removal strategy. In general, targeting 10% of the highest degree nodes resulted in the greatest decrease of the size of the largest component (GC) for all the hateful followers' networks; see Figure 6.5.

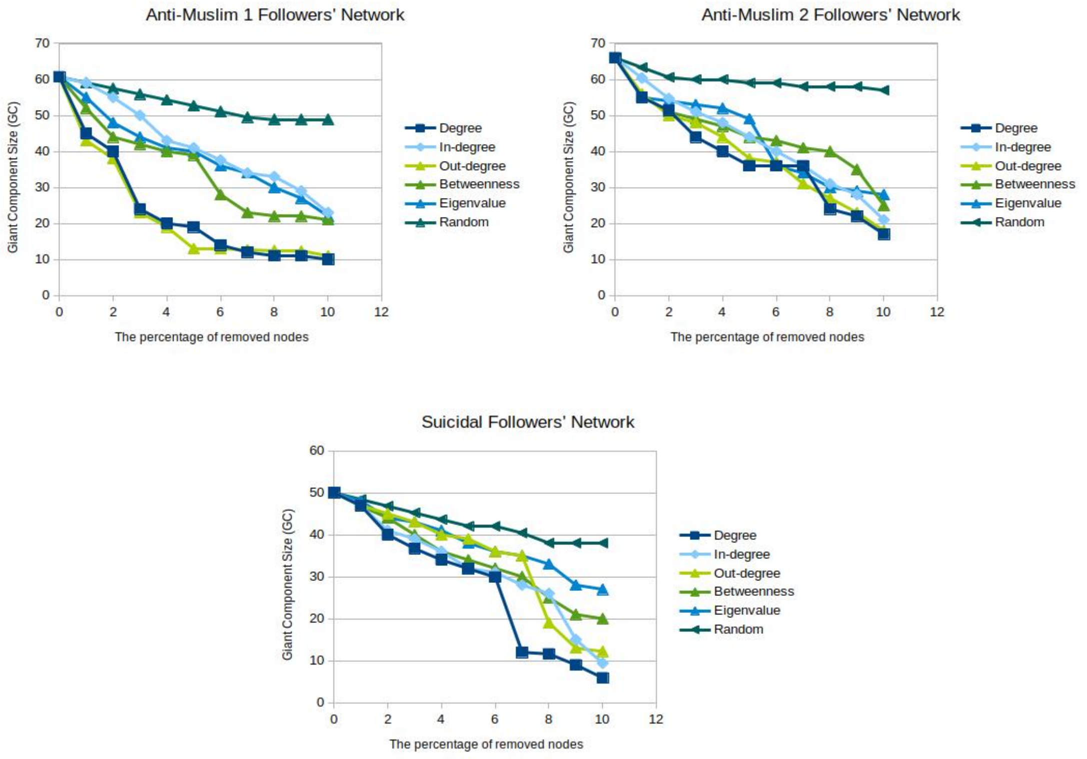


Figure 6.5: Impact of different removal strategies on the level of the Giant Component size of the hateful and suicidal followers' network.

For both hateful followers' networks, 75-83% of the largest component (GC) was disconnected, and this is also true for the suicidal network. In contrast, using other strategies, saw a reduction of only a 55-75% in GC size. From a network analysis perspective, the high degree nodes are usually connected to **'local bridges'** [143]. Local bridges are edges between two nodes in a community that are the shortest route by which information might travel from those connected to others in the same community [14]. In social networks, high-degree nodes also have a higher 'bridgeness' value as they have a higher chance of connecting to a **'global bridge'** [164]. Global bridges are edges between two nodes in two communities that are the shortest route by which information might travel from one community to another community [164]; in other words, the global bridge connects two different communities [4]. This suggests that the

[4] Note that local bridges differ from global bridges in that the endpoints of the local bridge once the

highest degree nodes in the hateful followers' networks have a 'bridgeness' characteristic (are more likely to be connected to global bridges). Thus, when these nodes were removed, the connected edges are also removed or changed consequently. Practically, the removal of users who are highly followed by others (influential) or who follow a lot of others (super-consumers), disconnects a community into two communities.

Moreover, it was observed that node removal based on out-degree (people who follow a lot of others) was more effective in terms of reducing the GC than in-degree for the hateful followers' networks. This is expected as the previous chapter suggested that both hateful followers' networks have fewer people with lots of followers, or 'influencers' (in-degree) and more people who follow a lot of hateful posters, or 'super-consumers' (out-degree). Here, we can see that those users who follow a lot of hateful posters are effective in connecting/disconnecting the hateful community compared to the influencers. This means that being influential in a hateful followers' network is not necessarily correlated with serving a crucial role for network connectivity - while the opposite was expected. This is a valuable finding for guiding policymakers not only to focus on targeting the influential and ignoring other 'effective' users.

Apart from the random-based strategy, the eigenvalue-based strategy was the least effective strategy for reducing the size of the GC for all the hateful networks, meaning that they are more robust in terms of removing those nodes which are connected with influential (high in-degree) or 'super-consumer' nodes (high out-degree). The eigenvalue-based strategy reduced the GC of Anti-Muslim 1 and 2 followers' networks by 63% and 57% which is less than the reduction achieved by a degree-based strategy. It is possible that the explanation for this is that nodes with high eigenvalues are more likely to be connected to important nodes (high degree nodes) by the local bridges rather than to the global bridges [155]. Granovetter *et al.* [135] stated that when the nodes connected to the local bridge are removed, the nodes on either side of the bridge become reachable only via very long paths. Thus, the eigenvalue-based strategy had

bridge has been deleted cannot have an edge directly between them and should not share any common neighbours.

less impact on reducing the size of the largest component because it resulted in a sparse largest component rather than disconnected components.

The degree-based and outdegree-based strategies also reduced the density of the hateful networks by 47-76% for Anti-Muslim 1 and 2, respectively, whereas other strategies recorded reduction not more than 43%, shown in Figure 6.6.

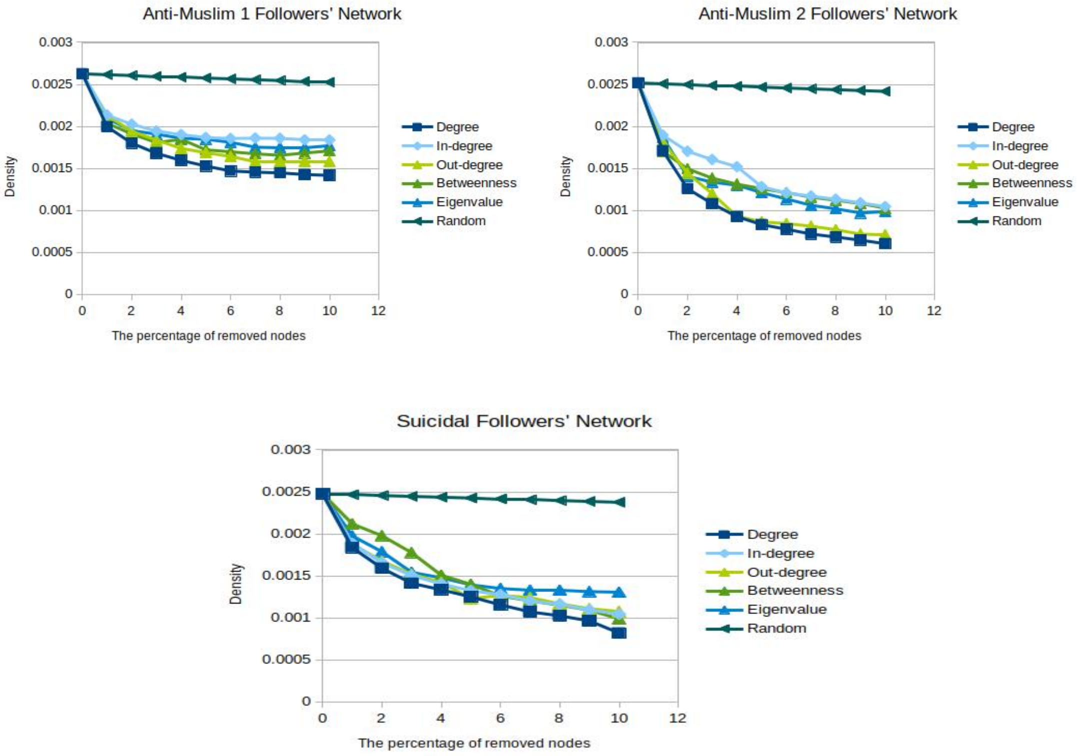


Figure 6.6: Impact of different removal strategies on the level of the density of the hateful and suicidal followers' network.

This is because removing the highly linked nodes (high degree nodes) reduces the number of connected edges in the network [206, 48]. Reducing the graph edge (links) means reducing the density, as the density is the ratio between the number of edges in the graph and the total number of possible edges. Less positive impact on the density reduction was observed when the betweenness-based strategy was applied to the hateful networks, compared to the degree-based strategy, despite the finding that the betweenness-based strategy was the most effective strategy for reducing the scale-free network connectivity in [40, 160]. This means that people who potentially allow hateful content to pass from one part of the network to the other are contributing less to the connectivity between the followers' networks users compared to high degree nodes.

Figure 6.7 illustrates that applying node removal strategies to the followers' networks shows fluctuating decreases and increases in average shortest path as nodes are gradually removed.

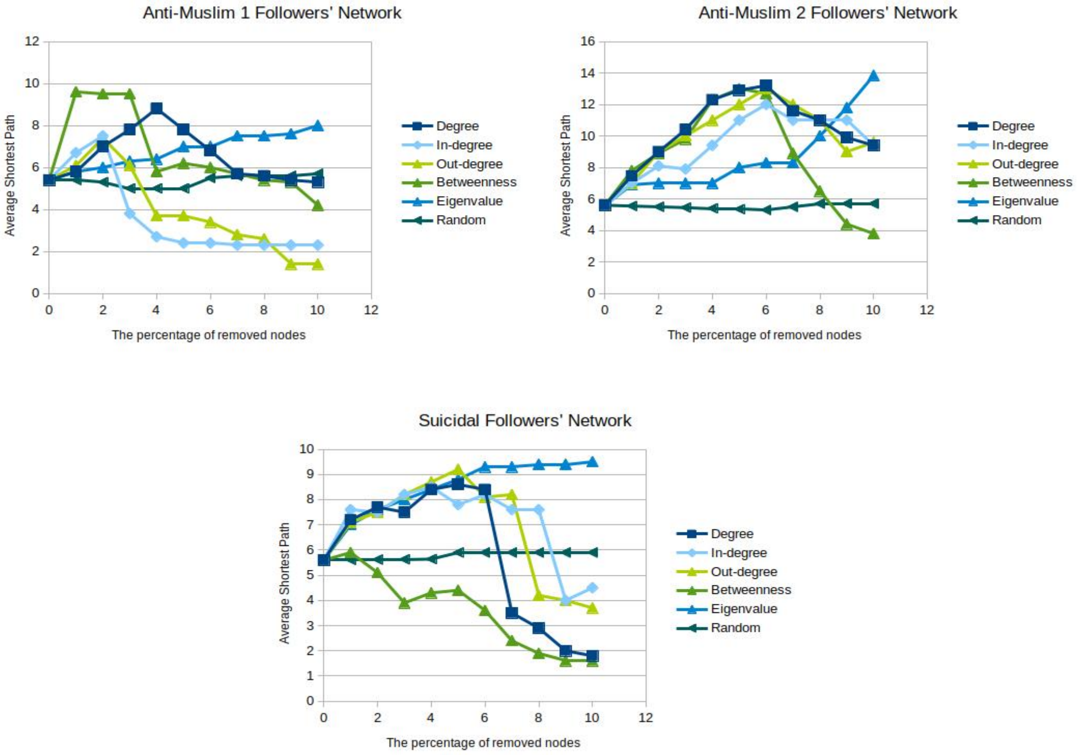


Figure 6.7: Impact of different removal strategies on the level of the average shortest path of the hateful and suicidal followers' network .

Noticeably, the eigenvalue-based strategy elongated the average shortest path of the Anti-Muslim 1 followers' network from 5.4 to 8, Anti-Muslim 2 from 5.6 to 13.8 and also from 5.05 to 9.5 for the suicidal network after removing 10% of the nodes. While the eigenvalue-based strategy had less impact on reducing the size of the largest component, it is possible that this strategy has resulted in a sparse largest component (not disconnected) and has therefore elongated the average shortest path. Elongating the average shortest path means that the average number of steps that are needed to deliver content to all the users in the biggest connected community is increased. In other studies, they refer to this effect as the *efficiency* measure. The efficiency measure is based on the shortest paths between two nodes, i.e. the minimum number of links used to travel from one node to another[121]. Efficiency decreases with an increase in the nodes' shortest paths; thus, more efficient networks have a small average shortest path. For example, in the Anti-Muslim 2 follower network, a hateful tweet would potentially reach all the nodes in the largest community (GC) within 5.6 steps. By removing 10% of nodes that are connected to highly linked nodes (high eigenvalue nodes), an increase was seen in the number the steps needed to reach the majority of the nodes in the largest community to 13 steps - essentially obstructing the information flow and, in other words, decreasing the network efficiency. This is also true for the Anti-Muslim 1 and suicidal networks. Users who are connected to highly linked nodes (high eigenvalue nodes) are not necessarily influencers (high degree users) but they may appear to be important for the gradual reduction of hateful flow (but not for disconnecting a network).

However, generally, the fluctuation of the increase and decrease of the average shortest path metric that has been noticed for the majority of the removal strategies has been also found in previous studies [40, 39] suggesting that the average shortest path metric might not be considered optimal for consistently measuring the connectivity performance of hateful followers' networks. Indeed, the average shortest path metric is widely used as an indicator for node removal efficiency in previous studies [168, 170, 327, 343]. This metric was also examined in the previous chapter in order to characterise hateful followers' and retweets' networks. It led to interesting results regarding the networks' connectivity for both followers' and retweets' networks, which motivated the author to consider it an indicator of the efficiency of removal strategies. Nevertheless, the author here suggests that this measure is not useful when networks have been significantly disconnected, which is in line with the finding of these studies [54, 53].

6.3.2 Retweet Networks

Figures 6.8, 6.9 and 6.10 show the six removal strategies that were applied to the retweets networks. It was observed that of the six removal strategies tested, the random removal strategy showed the least impact on reducing the size of GC, the density, and the average shortest path. In general, targeting the highest degree nodes resulted in the greatest decrease in the size of the largest component (GC) and the density for all the hateful retweets networks. Figure 6.8 shows that the degree-based strategy reduced the size of GC by 94%, 85% and 75% for Anti-Muslim 1, the antisemitic dataset and Anti-Muslim 2, respectively, while for the same percentage of nodes removed, other strategies led to a reduction of between 60 and 70% in GC size. It appears that the nodes with a high degree in the hateful retweets' networks also have 'bridgeness' characteristics, as explained previously in Section 6.3.1 when the removal of these nodes disconnect a community into two communities. In reality, this means that users with a high degree play a crucial role in hateful content propagation in the retweets' networks, suggesting targeting these users may separate the network into smaller communities and therefore decrease the network's connectivity. Targeting the highest out-degree users, 'super-retweeters', appears a more effective strategy for reducing the GC of two hateful retweets' networks (Anti-Muslim 1 and Anti-Semitic) compared to targeting the highest in-degree users, 'influential', meaning that the highest out-degree nodes may also have higher 'bridgeness' characteristics compared to in-degree nodes, This suggests that users who are retweeting other users are slightly more important in terms of disconnecting the hateful retweets' networks compared to users who are being retweeted a lot, 'influential'. According to the previous chapter's results in Section 5.3.2, the distribution of users showed that the hateful retweets' networks had more 'influencers' compared to 'super-retweeters'. Thus, it was expected that removing influential nodes (with high in-degree) would be more effective in terms of reducing the spread of hate; however, the findings suggest that removing the ’super-retweeters’ is actually more effective, indicating that targeting such people (e.g. by suspending them) would more likely reduce the spread of hate. This may guide policymakers not only to focus on targeting the influential and ignoring other ’effective’ users, who are spreading the hate - possibly via counter speech.

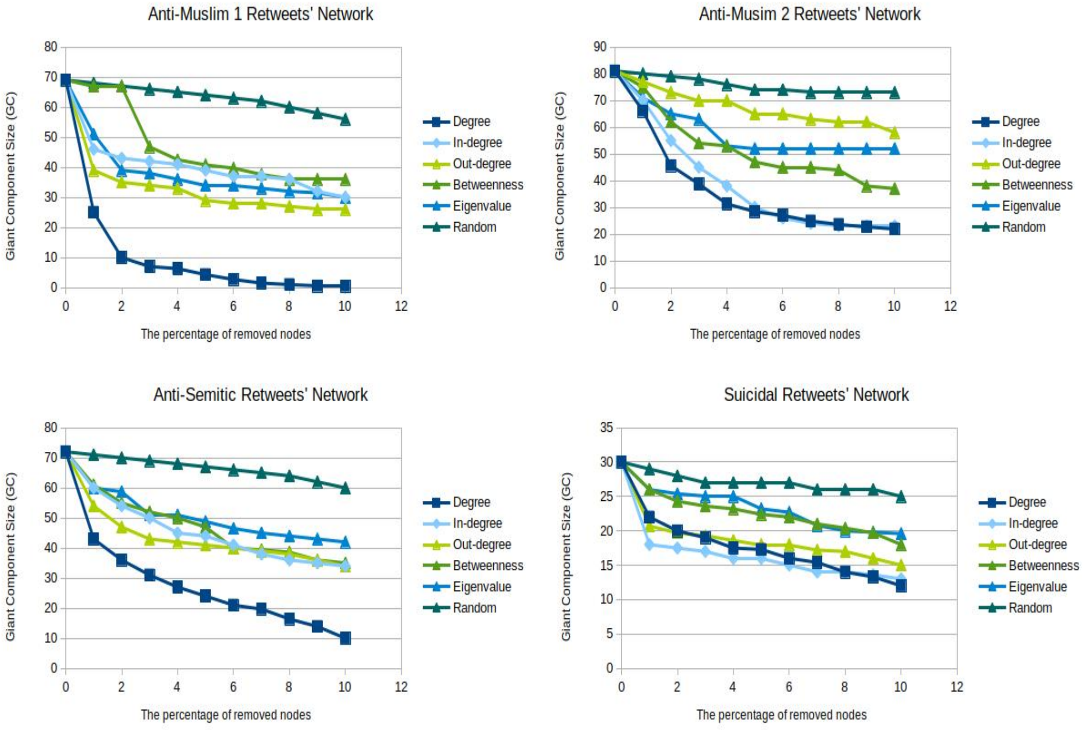
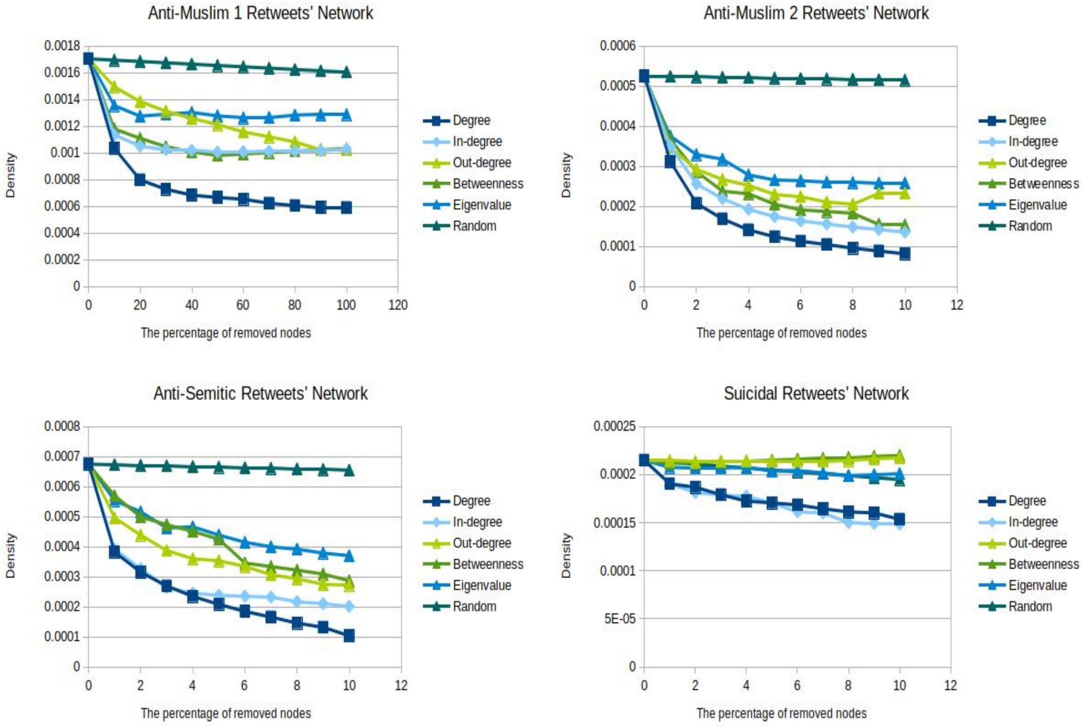


Figure 6.8: Impact of different removal strategies on the level of the Giant Component size of the hateful and suicidal retweets networks .

Moreover, Figure 6.9 shows that the degree-based strategy reduced the densities of the hateful networks by 65, 83 and 80% for Anti-Muslim 1, the antisemitic dataset and Anti-Muslim 2, respectively. This is also true for the suicidal network (indegree-based strategy is slightly more effective).


Anti-Muslim 1 Retweets' Network

Anti-Muslim 2 Retweets' Network

Anti-Semitic Retweets' Network

Suicidal Retweets' Network

Figure 6.9: Impact of different removal strategies on the level of the density of the hateful and suicidal retweets networks .

In contrast, other strategies recorded performances lower than the degree-based strategy by approximately 20%. This is in line with the results of previous studies showing that the most connected people (hubs) are the key players, being responsible for the greatest part of the spreading process [23][85]. Practically, removing the highly linked nodes (high degree nodes) reduces the number of connected edges in the network [206, 48]. Reducing the graph links decreases density as the density of real network systems (e.g. Twitter) is the ratio between the number of edges in the graph and the total number of possible edges. This suggests that in a propagation network, like retweets' networks, removing the highly linked users plays a key role in network vulnerability. The indegree-based strategy (removing influential users) appears slightly more effective in reducing the densities of hateful retweets' networks than the outdegree-based strategy ('super-retweeters'), while the latter has been shown in Figure 6.8 to be more effective in reducing the GC size than an indegree-based strategy. Reducing the density, using an indegree-based strategy, means decreasing the hateful users' links but not necessarily disconnecting the network's clusters (i.e reducing the GC size). This means that by removing the 'influential' (or high in-degree) users, the hateful links may be reduced; however, the GC may still be connected (the content is still able to flow among the clusters).

For the hateful retweets' networks, however, there is no single strategy that has a significant impact on increasing the average shortest path. Figure 6.10 shows a very fluctuated flow in the average shortest paths.

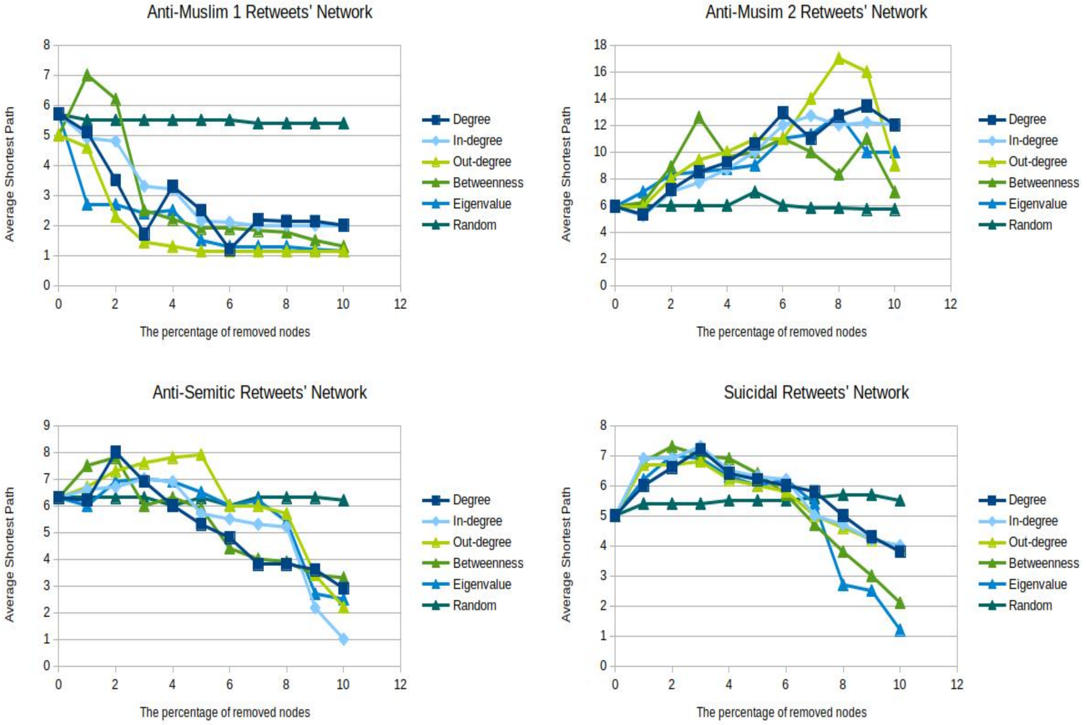


Figure 6.10: Impact of different removal strategies on the level of the average shortest path of the hateful and suicidal retweets networks .

Similar to the hateful followers' networks, it was also observed for the hateful retweets' networks that there is a fluctuation in the increase and decrease of the average shortest path metric among the majority of the removal strategies. This also suggests that this measure is not valuable when networks are significantly disconnected, which is in line with the finding of these studies [54, 53]. Next, we previewed the results of applying the hybrid node removal strategies (a combination of two different removal strategies) and investigated whether this combination would further increase the performance of the strategy and therefore decrease the network's connectivity.

6.3.3 Hybrid Node Removal Strategies

Followers Networks

Figures 6.11, 6.12 and 6.13 show seven hybrid removal strategies that were applied to followers' networks. In general, among the hybrid node-removal strategies, we noticed that the best removal strategies were the DegreeBetweenness and OutdegreeBetweenness strategies. However, Figure 6.5 shows that about 75-83% of the largest component (GC) was disconnected by the degree-based strategy while a lower percentage was obtained by the combined strategies, about 63-72% of the largest component (GC) was disconnected by the DegreeBetweenness-based and OutdegreeBetweenness-based strategies, respectively. However, the 'single' degree-based removal strategy in Figure 6.5 seems to be more effective in reducing the size of the largest component (GC) compared to the combined DegreeBetweenness/DegreeEigenvalue-based strategies in Figure 6.11. This means that removing nodes with the highest connection is the most efficient way to deconstruct the hateful networks; combining other strategies may reduce this effect. As mentioned previously, nodes with the highest connections have the characteristics of bridgeness, which is important for connecting the GC. Therefore, concentrating on this characteristic, by applying the degree-based strategy alone, is better than reducing its effect by combining it with another strategy; here the experiments ensure that.In addition, some 'single' node removal strategies performed better when combined with another strategy, and some are not (however, no combined strategy exceeded the performance of the 'single' degree-based strategy). For example, in the Anti-Muslim 1 followers' network, we noticed that combining the outdegree-based strategy with the betweenness-based strategy led to better performance than when applying the betweenness-based strategy alone but poorer performance than when applying the outdegree-based strategy alone. This suggests that the hybrid node-removal strategy may be advantageous when a targeted attack ought to focus on specific users (e.g. high betweenness users), so combining the outdegree-based strategy would strengthen the performance of the betweenness-based strategy.

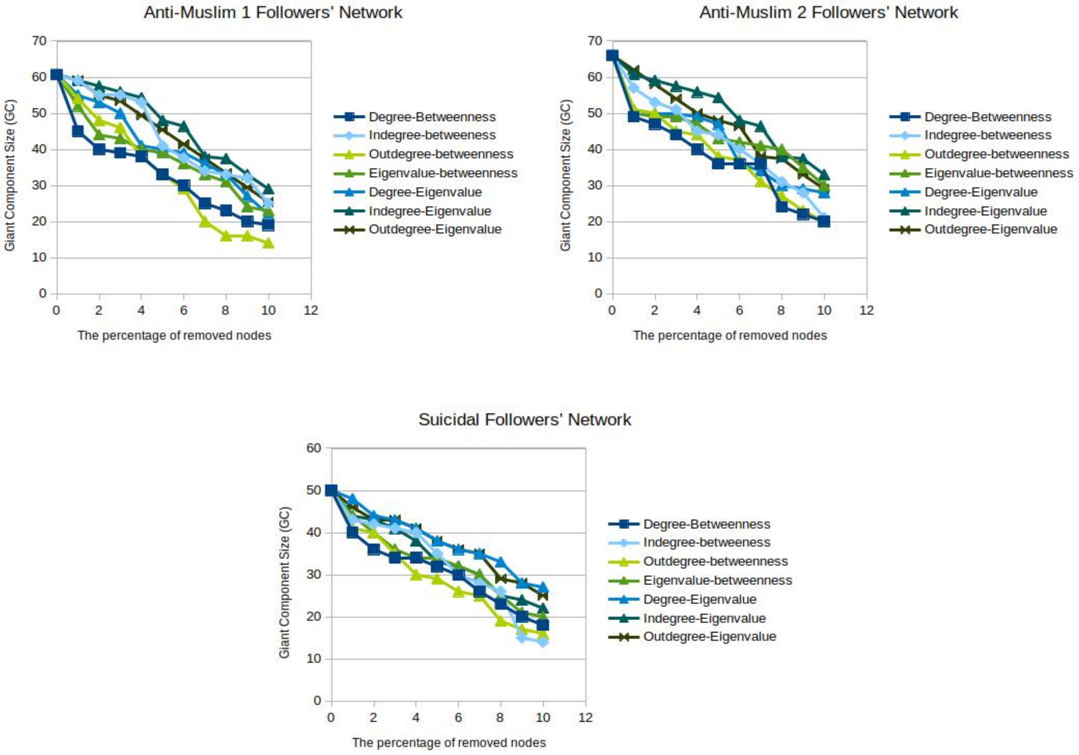


Figure 6.11: Impact of different hybrid removal strategies on the level of the Giant Component size of the hateful and suicidal followers' network..

Also, we noticed that the hybrid node removal strategies have less effect on reducing the density in Figure 6.12 and the average shortest path in Figure 6.13 compared to the 'single' node removal strategies. For density, the 'single' degree-based strategy removes the highly linked nodes (high degree nodes), which reduces the number of connected edges in the network and then reduces the density [206, 48]. The effectiveness of this was reduced when the single degree-based strategy was combined with another strategy. This suggests that the highly linked nodes (nodes with high degrees) are the node responsible for reducing the hateful networks links and therefore reducing the density as the density is the ratio between the number of edges in the graph and the total number of possible edges. For the average shortest path, the eigenvalue-based strategy lost its effectiveness on the gradual increase of the average shortest path when combined with other strategies. Additionally, we still observe a fluctuation in the increase and decrease in the average shortest path in Figure 6.13 which provides further evidence that the average shortest path metric might not be considered optimal for consistently measuring the connectivity performance of hateful followers' networks.

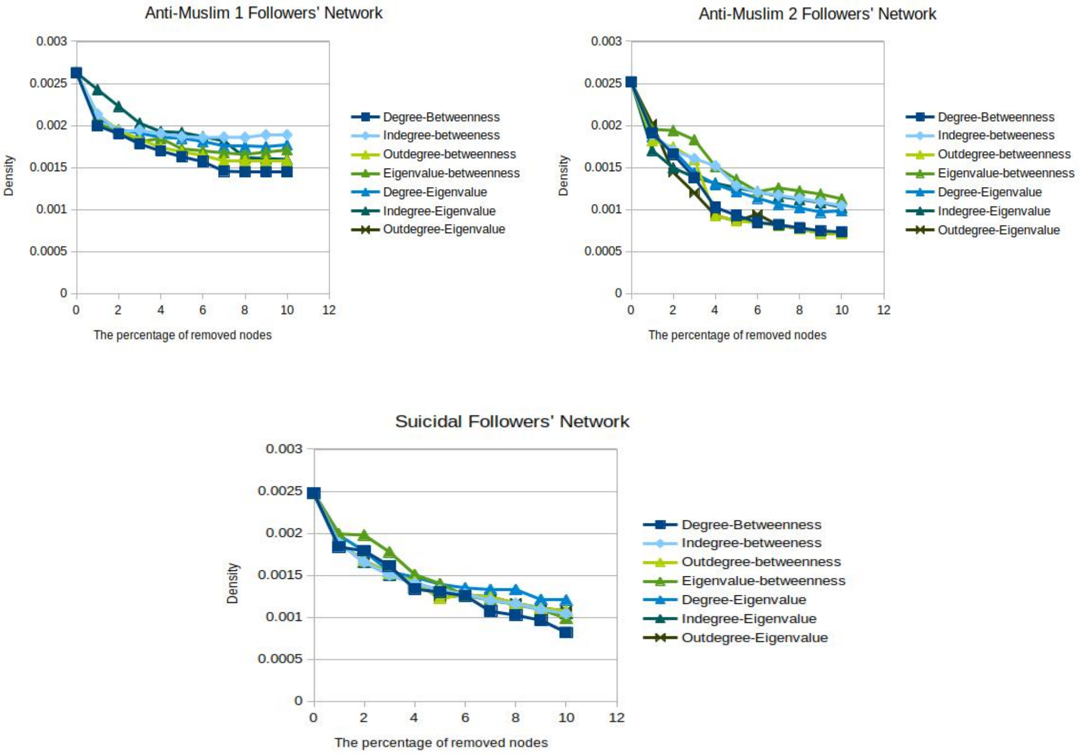


Figure 6.12: Impact of different hybrid removal strategies on the level of the density of the hateful and suicidal followers' network.

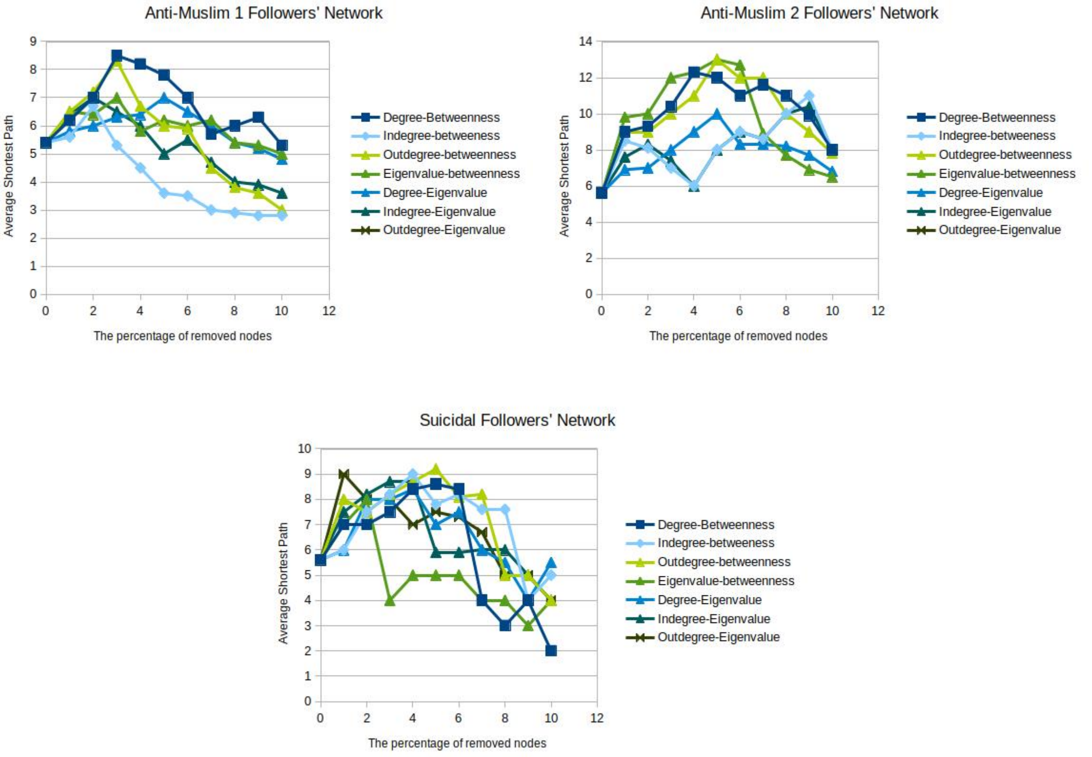


Figure 6.13: Impact of different hybrid removal strategies on the level of the average shortest path of the hateful and suicidal followers' network.

Retweets Networks

Figures 6.14, 6.15 and 6.16 illustrate the seven hybrid removal strategies that were applied to the retweets' networks. Similarly to what has been observed for the hateful followers' networks, all the hybrid removal strategies have less effect on reducing the size of the largest component (GC) and the density than the 'single' degree-based strategy. While the size of the largest component (GC) was reduced by 94%, 85% and 75% for Anti-Muslim 1, the antisemitic dataset and Anti-Muslim 2, respectively when applying the degree-based strategy alone, it was reduced by 86%, 63% and 69% by combining the degree-based strategy with other strategies. This is similar to what has been observed for the followers' networks, that combining another strategy with the degree-based strategy seems to reduce the advantage of the reduction in links obtained by applying that strategy alone as previously was shown in Figure 6.8. Here, we should emphasise the bridgeness characteristics for deconstructing the hateful networks.

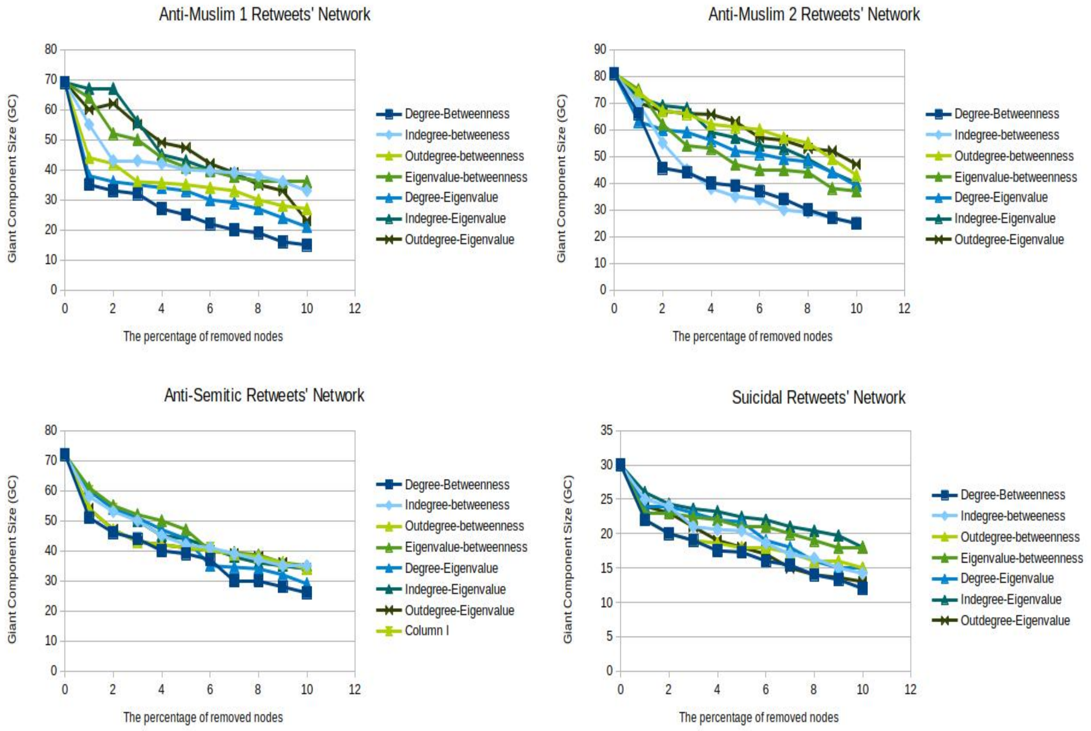


Figure 6.14: Impact of different hybrid removal strategies on the level of the Giant Component size of the hateful and suicidal retweets networks.

Moreover, for the density 6.15 we noticed that the combined strategies resulted in poorer performance compared to the ’single’ degree-based node-removal strategy in Figure 6.9. This suggests that the effectiveness of the degree-based strategy was reduced by combining other strategies with the degree-based strategy. Some ’single’ node removal strategies performed better when combined with another strategy, and some did not; however, no combined strategy exceeded the performance of the ’single’ degree-based strategy. Moreover, the fluctuation in the increase and decrease in the average shortest path in Figure 6.16 still exists, considering the average shortest path metric is not preferred for consistently measuring the connectivity performance of hateful retweets’ networks. This is also the case for the suicidal retweets’ network.

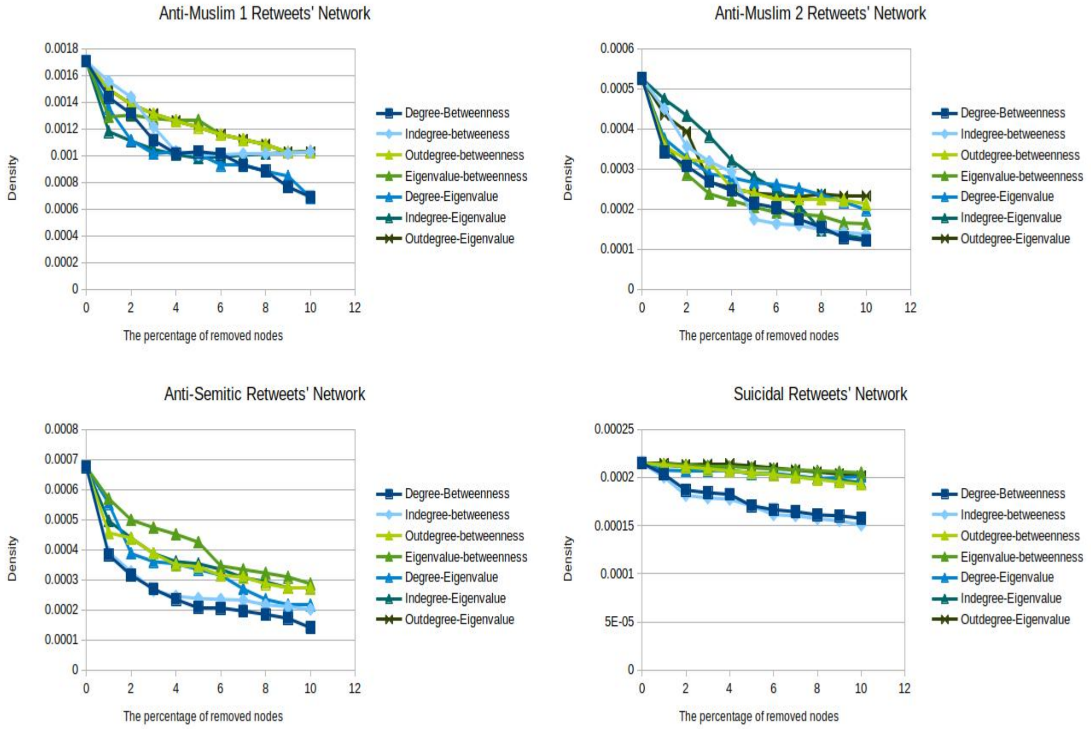


Figure 6.15: Impact of different hybrid removal strategies on the level of the density of the hateful and suicidal retweets networks.

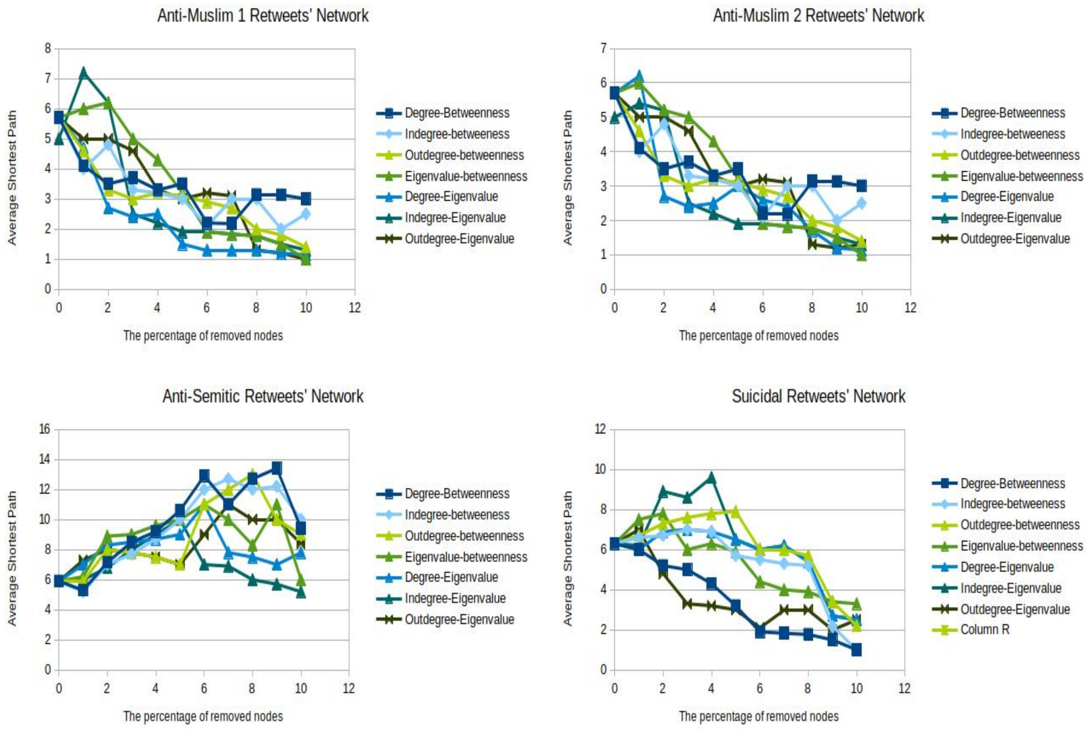


Figure 6.16: Impact of different hybrid removal strategies on the level of the average shortest path of the hateful and suicidal retweets networks.

So in general, the hybrid strategies seem to be less effective than the single strategies in this sample of the network. For example, it was expected that combining the degree-based strategy with the eigenvalue-based strategy would improve the results in terms of reducing the networks' connectivity (improve the results of the single degree-based and the eigenvalue-based strategies, separately). This is because the degree-based strategy was effective for reducing the largest component (GC), see Figure 6.5 and the eigenvalue-based strategy resulted in a gradual increase in the average shortest path among people in the followers' networks, see Section 6.3.1. Surprisingly, a combination of the degree-based strategy and the eigenvalue-based strategy did not decrease the networks' largest size nor did it increase the average shortest path compared to the effect obtained by the single degree-based strategy. This is in line with Bellingeri et al.[40] who showed that applying a combined attack strategy was the most efficient strategy for decreasing the largest component size (GC) in a scale-free network in the early stages of deletion (when deleting one or two critical hubs). Later in the attack sequence, the combined strategy was less efficient than the single strategy for attacking scale-free networks. This result has important implications for applied network science and deserves further investigation.

6.4 Removal Strategies on the Bipartite Networks

Figures 6.17 and 6.18 demonstrate the results of applying node removal strategies to bipartite hateful networks (followers and retweets).

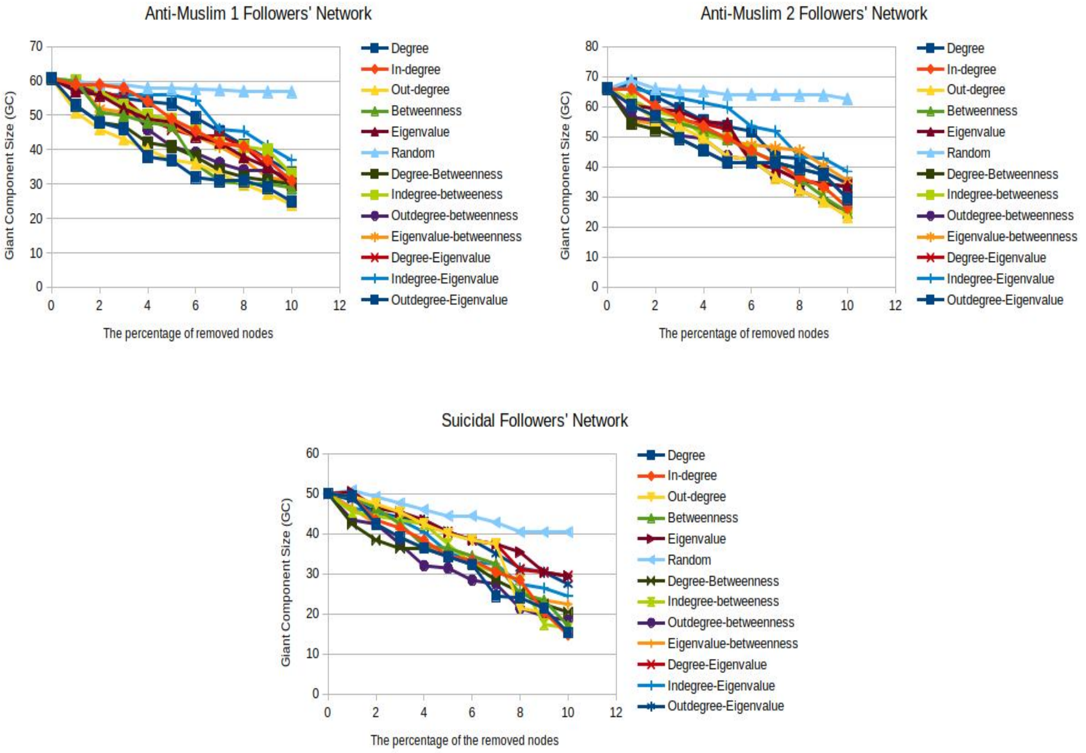


Figure 6.17: Impact of different node removal strategies on the level of the Giant Component size of the bipartite hateful and suicidal followers' network.

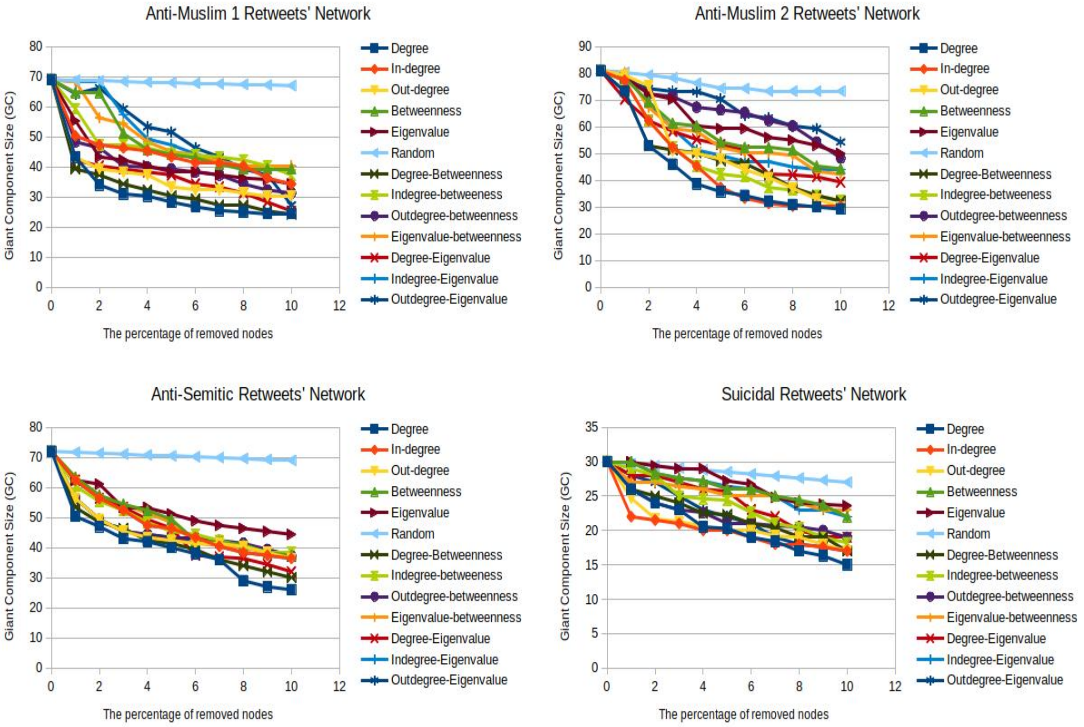


Figure 6.18: Impact of different removal strategies on the level of the Giant Component size of the Bipartite hateful and suicidal retweets networks.

First, we notice that no single or hybrid node removal strategy outperforms the single degree-based strategy on the generic network. The best performing strategy on the bipartite hateful followers' network is the outdegree-based strategy. However, it reduced the largest component by only 57-66%, significantly lower than the percentage obtained by applying the degree-based strategy to the generic hateful follower networks. Moreover, the best-performing strategy for the bipartite hateful retweets network is the degree-based strategy; however, its performance in terms of disconnecting the largest component is lower than its performance on the generic hateful retweets' networks. The degree-based strategy deconstructs 60%, 62% and 64% of the largest component for the Anti-Muslim 1, antisemitic dataset and Anti-Muslim 2 retweets bipartite networks, respectively. Meanwhile, the degree-based strategy reduced the size of GC by 94%, 85% and 75% for the generic Anti-Muslim 1, antisemitic and Anti-Muslim 2 generic networks, respectively.

Second, we observe that there is no significant difference among the node removal strategies themselves for the bipartite followers and retweets networks. This may be because the bipartite (two-mode) degree-measure has a higher correlation with eigenvalue and betweenness than the generic (one-mode) degree-measure [239].

There is no advantage in applying the node removal strategies of the bipartite networks to the generic networks which may give an indication of why those removal strategies that are applied to the generic network ('one-mode network') had a different impact on network connectivity compared to applying the same strategies to the bipartite networks [315, 122, 239]. Perhaps this is because of some differences in these networks' characteristics. First, the generic networks have one-degree distribution while the bipartite networks have two different degree distributions (top set/bottom set) [97]. So, instead of removing the highest degree nodes directly in the generic networks, in the bipartite networks, we remove the nodes by swapping between the bottom nodes, which represent the users and the top nodes which represent their affiliation (cliques). It has previously been mentioned that the overlap between cliques in the top nodes set is significant in the bipartite network, meaning that if two cliques have one node in common, then they certainly have many nodes in common [191]. This means that it is hard to disconnect two cliques because they are connected to each other by many nodes. Perhaps the bipartite version of the hateful networks in this sample contain a lot of overlapped cliques that maintain the node's connectivity. As a strategy for cutting off the networks that spread hatred, it could be beneficial in future to concentrate on how to target the overlapped cliques of the bipartite version of the hateful networks. Delete a node, for instance, along with its first, second, and so on neighbours. But it's important to keep in mind that the study's hateful networks are not inherently bipartite. Therefore, it may be fundamentally necessary to evaluate the significance of overlapping groups in this datasets' study.

Moreover, the projection of the bipartite network (converting the two-mode network into a one-mode network to compare the GC of the bipartite network with the generic networks) may have resulted in many highly connected nodes [353], as the remaining nodes with high degree induce large cliques compared to the generic version of the network. Also, the projected bipartite network resulted in high clustering coefficients (high clustering coefficients in a projection may be seen as a consequence of the underlying bipartite projection rather than a specific property of the network [190]) and likewise, the projection may lead to very dense networks compared to the generic version, even if the bipartite version is not dense. This phenomenon is illustrated particularly in [190, 246]. These observations may confuse the results.

Therefore, when performing SNA one must be aware of the characteristics of the projected network to avoid bias when comparing it to the generic version of that network.

To conclude, it is clear that we were effectively able to deconstruct the most significant component of the hateful followers and retweets networks by targeting the highest degree users of the generic networks (one-mode networks). However, the efficiency of node-removal strategies depends on the topology of the network [40], meaning that further investigation of different hateful networks is required. Table 6.3 summarises the results of the node removal strategies.

Table 6.3: Summary of the node removal strategies on the normal (one-mode) networks and bipartite (two-mode) networks.

	Node removal strategies	
	Single	**Hybrid**
Normal (one-mode)	1- Degree-based strategy. 2- 75-83% of (GC) was disconnected in the hateful followers' network. 3- 94%, 85% and 75% of (GC) was disconnected in the hateful retweets' network.	1- OutdegreeBetwenness-based and DegreeBetweenness-based 2- 63-72% of (GC) was disconnected in the hateful followers' network. 3- 86%, 63% and 69% of (GC) was disconnected in the hateful retweets' network.
Bipartite(two-mode)	1- Outdegree-based strategy. 2- 57-66% of (GC) was disconnected in the hateful followers' network. 3- 60%, 62% and 64%of (GC) was disconnected by Degree-based strategy in the hateful retweets' network.	DegreeBetweenness-based strategy and OutdegreeBetweenness strategy performed the best among the hybrid strategies. 57%-66% of (GC) was connected. However, not better than 'single' Degree-based strategy where 75%-94% of (GC) was connected. .

Table 6.3 demonstrates that the best-performing strategy in terms of reducing the (GC) size is removing the users with the highest degree of centrality on the generic followers and retweets networks. Combining two node removal strategies reduces the effectiveness of the 'single' node removal strategy. In addition, the bipartite networks have not shown much promise in terms of detecting the most central users compared to generic networks. However, we may still need to examine the node removal strategies for networks with more than two-modes, e.g. tripartite. These are referred to as N-partite networks or multipartite networks. An N-partite graph is a graph whose nodes are or can be partitioned into *n* different independent sets - forming such a partition that no two nodes belonging to the same subset are adjacent. The overlap among the network cliques becomes greater in a multipartite network [191]. Despite the fact that the overlap among the cliques of the bipartite networks has not improved the performance of the node-removal strategies in our experiments, it would be interesting to examine the node removal strategy on networks with a higher intersection among the cliques (multipartite).

6.5 Conclusion

This chapter investigated strategies for disrupting hateful networks. For this investigation, a range of node-removal strategies targeting users depending on their role in the network was developed. A simulation of 13 node-removal strategies was applied to generic networks and a bipartite version of the generic networks indicated that targeting nodes with the highest degree in the generic hateful networks is the most effective way of reducing the largest number of hateful followers' and retweets' networks compared to the other strategies, thus limiting the exposure and transmission of hateful content.

Targeting nodes that highly follow or retweet other hateful users (outdegree-based strategy) also had a significant effect on reducing the largest component and essentially obstructing the hateful information flow (except for the Anti-Muslim 2 retweet network), suggesting that nodes with higher out-degree - the 'super-consumers' and 'super-retweeters' of the hateful content - constitute indispensable bridges responsible for connecting different clusters in a network. It could be assumed that removing influential nodes (with high in-degree), would be more effective in terms of reducing hate exposure and spread. In fact, it seems that being influential in a hateful network is not necessarily correlated with serving a crucial role for network connectivity. However, the findings suggest that removing the 'super-consumers/super-retweeters' (high out-degree) is actually more effective, indicating that targeting such people (e.g. by engaging them either by suspension or counter speech) would more likely reduce the spread of hate. This is a valuable finding for guiding policymakers not only to focus on targeting the influential and ignoring other 'effective' users.

The experiment also showed that the combination of two node-removal strategies (hybrid) has not provided any further effect on reducing networks' connectivity compared to applying the 'single' degree-based strategy. Moreover, we noticed that there is no advantage in applying the node-removal strategies on the bipartite networks compared to the generic networks. The small sample size in this study may have an impact on this finding, which indicates a potential and/or hurdles to generalising such findings. As far as the author knows, there aren't enough research that specifically examine the relationship between dataset size and node removal tactics. Regarding the impact of the dataset size on the node removal strategies, more research is required.

Broadly speaking, social networks are resistant to node removal, and the results of the strategies were not inevitable. In addition, it is notable that some strategies are effective on specific metrics, and some are not. By way of an example, the degree-based strategy was the most effective strategy for disconnecting the hateful networks, while the eigenvalue-based strategy may be effective for gradual elongation of the average shortest path of a hateful network. This suggests, for a future work, that combining different node removal strategies might have a significant impact on reducing the entire network's connectivity and flow.

In the end, we should note that the findings of this study only relate to the sample at hand, and may not be generalisable beyond this study.This is a limitation that was also recognised in the previous chapter, that the data examined for this work are still likely to represent a sub-section of the entire hateful networks and hateful users engaged in posting hateful tweets, suggesting that these findings may be applicable to a small sample for a hateful network. However, obtaining the entire network structure is difficult as real-world networks such as social networks and the Internet are massive and typically only part of the network structure is available from network [132, 211, 218].

Chapter 7

Conclusion

7.1 Introduction

This final chapter provides a summary of the research conducted in the thesis. After summarising the main contributions of the thesis, it describes the key observations related to the research questions and contributions to the field. Then follows a discussion of the implications and limitations of the thesis. Finally, it highlights some possible directions for future work.

7.2 Thesis Summary

Researchers have concentrated on exploring online hate content on specific social media platforms, e.g. Twitter [63, 224] and/or using computational methodologies to detect, understand, manage and remove hateful content [229]. However, there are issues remaining that require attention that relate to the methods of online hate detection and management. The motivation of this thesis was to contribute to the growing literature in terms of measuring online hate, focusing particularly on detecting, characterising and managing the phenomenon.

Specifically, the main aims of this thesis were as follow: (i) to improve contextual cyberhate classification; (ii) to characterise multiple hateful networks in terms of exposure to and the propagation of cyberhate; and (iii) to demonstrate how to disrupt or manage cyberhate propagation on Twitter. Together, these form a cohesive proposition for the management of cyberhate - and are the core contributions of this thesis as a whole.

Chapter 2 introduced comprehensive knowledge relating to the definition of cyberhate, and summarised the literature related to the thesis work.

Chapter 3, provides an overview of the proposed methods for hateful content classification, hateful networks characterisation and hateful networks disruption.

In order to improve the contextual cyberhate classification, Chapter 4 presented work on the classification of text that does not contain clearly hateful words which would have an impact on classification accuracy. It proposed a novel combination of semantic learning and psychological theory for improving machine learning for automated cyberhate detection.

Regarding the characterisation of hateful networks, Chapter 5 aimed to understand hateful networks' connectivity in terms of the exposure to and the propagation of cyberhate. This baseline study characterises multiple hateful networks extensively. Investigating such characteristics is required in order to better understand hateful networks, which is considered an important step prior to controlling, detecting or disrupting propagation.

As a consequence, Chapter 6 aimed to find the best strategies for disrupting hate exposure and propagation by applying a wide range of node removal strategies and then, measuring the changes in network connectivity post node removal to assess whether the structure of the network significantly changed and subsequently 'disconnected' users.

Overall, the research conducted for this thesis, mainly Chapters 4, 5 and 6, has made significant advances in the field of measuring, understanding, and disrupting the propagation of cyberhate on Twitter.

7.2.1 Thesis contributions and key observations

Chapter 2, refined our understanding of hate speech and provided an insight into the wider research area, which provides the basis for shaping the contributions of this thesis (surrounding hate speech classification, propagation and disruption methods), and highlighted the limitations. Moreover, the discussion on the existing literature and on identifying research gaps led to open research questions. The research questions identified in Chapter 2, helped shape the structure of the thesis. In the next section, each research question is repeated, and the relevant contribution discussed, including any related analysis and new knowledge that has been acquired.

Cyberhate Classification

There have been numerous attempts to automatically classify, identify, and quantify cyberhate. This has included lexicon [131, 180, 153], syntactic [136, 243] and semantic [183, 36, 106] features. However, these approaches are limited as their classification of text does not include words which have clear hatefully antagonistic content (for example *'send them home'*), which would have an effect on overall classification accuracy. Chapter 4 presented a unique method for classifying cyberhate based on the use of 'othering language', and drew upon Intergroup Threat Theory (ITT). In particular, it determined whether using pronouns referring to an ingroup (i.e. 'we', 'us') coincided with pronouns referring to an outgroup (i.e. 'them', 'they'). This process indicated the use of divisive or antagonistic language and can subsequently improve machine classification of cyberhate. This addressed the following question:

RQ 1: ***To what extent can using othering and ITT theories to drive the development of new features for classifying cyberhate improve the performance of machine learning for cyberhate detection?.***

This was used to build a feature set that is referred to as an 'othering feature set', which is utilised to enrich the representation of text examples of cyberhate. These features are subsequently employed in combination with a paragraph embedding algorithm that infers semantic similarity between features to create a model that represents 'othering' language which is used for the purposes of cyberhate classification. The experimental results in Chapter 4 showed the efficacy of including the feature set for classifier training over the baselines model, and when trained using 10-fold cross validation produces a 0.99 F-measure for all three models. The results are an improvement upon previous findings in the cyberhate literature, which provides evidence that the othering narrative could play a significant role in detecting hate speech.

The three models which were determined to be the best performing were then tested on different unseen datasets using four different variations of cyberhate. These models were:

- The comprehensive classifier. This was an implementation of the work of Nobata *et al.* [243] combined with our novel othering feature set, and the raw dataset. It was termed comprehensive as a broad range of features were applied, including n-grams, linguistics, syntactic words and comment embedding.
- The othering + raw classifier. This was based on the implementation of Paragraph2Vec feature extraction and an MLP classifier, using our novel othering feature set and raw dataset.
- The othering classifier. This refers to the proposed application of Paragraph2Vec feature extraction and the MLP classifier on the proposed othering feature set (without raw text).

For the four dataset types - religion, disability, race and sexual orientation - F-measures of 0.81, 0.71, 0.89 and 0.72 respectively were obtained on unseen data. However, these 'optimum' results were not obtained by a single model. Different models resulted in different performances for various types of hate. The othering + raw classifier exhibited the lowest FNs (missed instances of hateful samples) for religious and racial hate, which suggests that the othering language features could be used to detect cyberhate towards religious and racial groups. The comprehensive classifier includes the additional n-gram features in addition to the othering language, which performed most effectively in the detection of disability and sexual orientation hate. Solely using othering language features subsequently exhibited less effectiveness in the detection of hostility towards these groups. The additional text features, such as n-grams and linguistics, were better at capturing the hate features directed at these groups, indicating that different types of hate speech have different language characteristics. The take away finding is that, although using othering terms is effective for some contexts of hate speech, it is not effective for all. The experiments revealed that othering language was a valuable feature in the identification of anti-religious and racial content, although it had less value in the context of disability and sexual-orientation hate.

The implementation of the novel hate speech classification process generated the first contribution of the thesis:

C1:*The development of a novel 'othering' feature set that utilises language use around the concept of 'othering' and Intergroup Threat Theory to identify the subtleties of implicit cyberhate. A wide range of classification methods was implemented using embedding learning to compute semantic distances between parts of speech considered to be part of an 'othering' narrative and improve the state-of-the-art embedding models in cyberhate detection by 2%-59%. When tested on unseen data using four different types of cyberhate, namely: religion; disability; race and sexual orientation, F-measures of 0.81, 0.71, 0.89 and 0.72 were obtained, respectively. Furthermore, the experiments show that different types of hate speech have different language characteristics and the use of othering terms can be effective for some but not all contexts of hate speech.*

In fact, previous studies had very little focus on classifying text that did not contain hateful words (implicit hate) which would have an impact on classification accuracy, *(e.g. get them out)*. Burnap *et. al.* [63] suggested that utilising psychological theories for detecting cyberhate would contribute to improving the detection process. Indeed, this researcher's method was to interpret certain statistically effective linguistic features though they did not test these features with machine classification algorithms and state-of-the-art features such as semantic features (that may capture the similarity between hateful terms). Building on that work, our first contribution uses the 'othering' language for detecting one sort of implicit hate speech - the language that distances other groups. Although implicit language is not limited to 'othering' language, identifying it contributes to detecting some implicit hate, improving the classifier performance. This is achieved by reducing the reliance on lexical features, which helps alleviate overfitting to the training dataset. Also, domain knowledge such as linguistic patterns and underlying sentiment of hate speech can inform model design, feature extraction or preprocessing. In addition, despite the fact that the previous methods reported promising results, it is apparent that their evaluations were not generalised on an external dataset. Our experiments were tested on four unseen datasets relating to four types of hate speech. This sort of testing can be a valuable tool for measuring generalisability; we need generalisability in the real world.

However, our approach presents some limitations; first, we used a considerably small sample size dataset for training and testing our classifier, while it is preferable to implement the embedding learning on a large amount of text data to ensure that valuable embeddings are learned. This is due to the limited number of relevant datasets that are publicly shared and also the high cost of human annotation. Second, by testing our model, we assumed that tweets which contain two-sided othering language (othering pattern) are more likely to be hateful. However, this is not necessarily the case. For example, a tweet like 'send them' may imply hate if it appears in relation a hateful event, but neutral when it appears in a marketing-related tweet. The point here is that implicit cyberhate needs more studies to investigate its patterns.

Cyberhate Network Characterisation

Social media allows individuals and groups to disseminate ideologies, which at times can be detrimental to societal cohesion and individual well-being, for instance, online hate. It has consequently become important to examine the online structure, communication, and connectivity of online communities to ascertain the exposure of users to hateful ideologies which may influence their own actions and opinions, or lead to online harms. Detecting online hate speech has been extensively examined from a content analysis perspective (including the previous contribution in this thesis). However, research into characterising hateful networks on social media is not well examined in the literature.

Some research has applied social network analysis (SNA) to Twitter in order to use connectivity information to indicate when a user has posted offensive content [273, 19]. Other studies have focused SNA analysis on retweets to measure diffusion [278, 273]. However, no research has yet examined multiple networks to ascertain if there is evidence of similar levels of friendship and therefore a general exposure to the hate, or similar levels of propagation behaviour and therefore a general contagion effect. The lack of a Twitter study led to the following research questions:

RQ2: By studying multiple hateful networks on Twitter, is there evidence of similar of 'levels of friendship' across multiple hateful networks, and therefore a general measure of exposure to cyberhate?

RQ3: By studying multiple hateful networks on Twitter, is there evidence of similar levels of propagation behaviour and therefore a general contagion effect?

Chapter 5 described a baseline study extensively characterising a number of hateful networks from different perspectives in order to expose cyberhate (in the follower networks) and propagate cyberhate (in the retweets network). The values of various SNA metrics were utilised to answer questions that increased the understanding of the connectivity of the hateful networks.

With regard to exposure to hateful content, Chapter 5 described the collection and analysis of two anti-Muslim followers' networks. The objective was to discover if there were any similarities between these networks and compare them with another 'risky' network with similar characteristics, i.e. a suicidal follower network. The findings revealed that users within the hateful follower networks were at a similar level of exposure to hateful content; by comparison, users within suicide networks were less likely to be exposed to risky content, by around 10%. The size of the largest component and density was marginally higher for both hateful follower networks than the suicide network by over 10%, and 0.0002 for the largest component and density, respectively. This indicates a higher exposure to, and potential virality of, hateful content in the sample. It was also found that both hateful networks had fewer 'influencers' (people with many followers), and more 'super-consumers' (those following many hateful posters), while the suicidal follower network displayed a slightly higher volume of influencers, and less nodes following a large number of similar users. We interpreted this as indicating that hateful follower networks in the sample had a tendency to be more vulnerable to content exposure [199].

In addition, the clustering coefficient metric showed that the probability of neighbouring nodes which are also connected (densely connected neighbours) was consistently higher for the hateful networks than the suicidal network. Both cases had densely connected neighbours, irrespective of whether the nodes were hate 'super-consumers' (out-degree edges) or hate 'influencers' (in-degree edges). The behaviour was slightly lower in the suicidal follower network, indicating greater connectivity between the hateful users, and therefore increasing the risk of hateful content exposure.

A comparison of the average shortest path of the follower networks with the relative largest component size indicates that the hateful follower networks had greater connectivity than the suicidal follower network. Five steps were required to reach over 60% of the users of the hateful followers networks, whilst five steps were only able to reach half of the suicidal users. Moreover, the average shortest path of the hateful networks was similar to a more extensive Twitter network, which indicated that the hateful follower networks exhibited data flow properties which resemble larger scale communication follower networks, although in a very small-scale network. Yet, they had less reciprocated friendship behaviour than the suicidal users (less connected around the topic). These experiments and associated observations gave rise to the second contribution of the thesis:

C2: *To the best of the author's knowledge this is the first study carried out to understand the connectivity characteristics of two hateful follower networks. The analysis shows that the level of connectivity of the hateful followers' networks is similar, and therefore results in common levels of users' exposure to cyberhate. Hateful networks were also compared to another form of 'risky' network (i.e a suicidal ideation network of similar size) to understand the general level of the hateful networks' connectivity. The results showed evidence of higher connectivity between the hateful users (higher exposure to hateful content) compared to the suicidal users, which suggests a potential virality of hateful content. They, however, have less reciprocated friendship behaviour than suicidal users (they are less connected around the topic). In addition to the contribution of the knowledge, this study resulted in the first friendship datasets in the field that could be used in further research studies.*

With regard to the propagation of hateful content, Chapter 5 also collected and analysed three anti-religious retweets' networks with the objective of discovering any similarities that exist between these networks, and subsequently comparing them to another

’risky’ network on suicidal retweets. Several structural similarities were found, including differences between the hateful retweet network in terms of social network metrics. The largest component of hateful retweet networks was greater than 69%, with densities greater than 0.0005, and reciprocities greater than 12%. Every metric is higher than those recorded for the suicidal retweets network. This shows that the hateful retweeters in the sample exhibited consistently high levels of information propagation behaviour, with increased content contagion in the hateful retweet networks compared to those in the ’risky’ network.

Moreover, the degree distribution of the hateful retweets’ networks showed that popular users (high in-degree) are responsible for creating information cascades as they are highly retweeted by other hateful users. Moreover, there was a considerable number of hateful users (high out-degree) engaged significantly with hateful conversation by retweeting other users’ hateful messages (but less than high in-degree), while the suicidal network did not exhibit this behaviour. This suggests that more co-operation on the spread of the message (hate) in hateful networks, across all three hateful networks, and less in the suicide network would be seen.

Moreover, comparing the average shortest path value with the largest component indicates that the hateful retweet networks required from 5-6 steps to reach 70-80% of the users, while 5 steps only reached 30% of the suicidal users. This shows high reachability (contagion) of the propagated messages amongst the hateful users, meaning that while an identical number of steps are required to reach the largest cluster in the sample, the hateful content reach is consistently greater than the suicide content - with 37.9-50% more users reached by retweeting.

With regard to reciprocity, the hateful retweets networks exhibited significantly higher reciprocity than the suicidal retweets network, indicating that greater co-operation on spreading messages in hateful networks can consistently be found across all three hateful networks in the sample, and less in the suicide network [291]. However, there should be a wariness around this finding, as the high reciprocity level amongst hateful retweets networks may assist with building a collaborative network which ultimately increases the co-operative levels of hate propagation. The density and average clustering coefficients were not indicative of significant smaller 'organised' subgroup exits (where users all retweet each other). However, there was consistent reciprocal retweeting with regard to hateful content which seemed to be higher than a comparator network. These analyses and related observations lead to the third contribution:

C3: *To the best of the author's knowledge this is the first study carried out to understand the communication characteristics of three hateful retweets networks. Analysis shows several structural similarities were observed among the retweets' networks as were differences between the hateful retweet network. Also, there was a consistently and significantly greater reach of content (contagion), and greater degree of co-operation on the spread of the message (hate) in hateful networks, across all three hateful networks, and less in the comparator 'risky' network - suicidal ideation. Hateful content reaches more users in fewer hops.*

The current literature on characterising hateful Twitter networks has focused on the propagation of tweets (retweets), not on the tweets' exposure (followers). Moreover, the characterisation process in the literature was implemented on one network while we implemented our characterisation on multiple networks in order to be able to generalise the results. Table 2.6 in Chapter 2, demonstrated that there are three studies related to characterising hateful networks on Twitter; however, they characterise only the propagation (tweet networks) of a single hateful network while this thesis touched on hateful network exposure follower networks). In addition, the followers' networks are the first in this field. These networks are not straightforward collections. It was time-consuming to collect and build them containing only the hateful users. The publication of these networks will contribute to updating the literature as no similar networks have been built and published. However, generalising these results needs further research with more extensive networks. Indeed, this study was conducted on a limited-sized set of annotated posts. Thus, it is still likely to represent a sub-section of the entire sample of hateful networks and hateful users engaged in posting hateful tweets, suggesting that our findings may be applicable to a small sample. However, because real-world networks, such as social networks and the Internet, are huge, obtaining the entire network structure is difficult, and typically only part of the network structure is available from a network [132, 211, 218].

Cyberhate Disruption

It has been indicated that exposure to and engagement with cyberhate is linked (although possibly not causally) to online hate production and even offline aggression, as some perpetrators of violent hate crimes are known to have engaged with such content [102]. Regulation of social media platforms is a challenge [177], and policymakers have had difficulty determining practical ways to reduce cyberhate [7]. Attempts to ban and remove hate-related content have not proven to be effective [1] because of concerns about limiting freedom of expression. Therefore, controlling or disrupting the flow of cyberhate is required, in addition to reducing the propagation of, and exposure to, hateful content by removing specific users (nodes) and therefore disrupting the flow of hate. At the time of writing the literature, it appeared there was no research which had applied disruption strategies to Twitter hateful networks in an attempt to reduce network connectivity (exposure reduction) and therefore diffuse hate (contagion reduction). Chapter 6 attempted to provide an answer to the question:

RQ4: According to the networks' structural characteristics, which node removal strategies would be most effective at decreasing the potential exposure to or the propagation of hateful content?

In Chapter 6, we experimented with various disruption methods - or node-removal strategies - to reduce the propagation of cyberhate on Twitter through network pruning. Thus, various node-removal strategies were developed to target users depending on their characteristics. A simulation of six node-removal strategies showed that it is more effective to target nodes with the highest degree, and that this strategy can more effectively reduce the largest component of hateful followers' and retweets' networks compared to the other strategies, which limit the exposure and transmission of hateful content. For both hateful followers' networks and the suicide network, 75%-83% of the largest component (GC) was disconnected. Conversely, other strategies resulted in a reduction of only 55%-75%. For the retweets' networks, the degree-based strategy reduced GC size by 94%, 85% and 75% for Anti-Muslim 1, the antisemitic dataset, and Anti-Muslim 2, respectively; while the same fractions of nodes led to a reduction of 60%-70% in GC size. In addition, the degree-based strategy reduced hateful network density by 65%, 83% and 80%, respectively. Other strategies recorded performances approximately 20% lower than the degree-based strategy.

More specifically, it was also found that targeting nodes which retweet other hateful users, or follow them to a high degree, also significantly reduced the largest component and obstructed the flow of hateful information (except for the Anti-Muslim 2 retweets' network). This suggested that nodes with higher out-degree - the 'super-consumers' and 'super-retweeters' of the hateful content - bridge the connection of different clusters in a network. There may be an assumption that removing 'influential' nodes (with high in-degree), would more effectively reduce the spread; however, the findings indicated that removal of the 'super-consumer' and 'super-retweeters' was actually more effective, suggesting that targeting these people (e.g. by suspending them) would potentially reduce the spread of hate.

Users who are connected to highly linked nodes (high eigenvalue nodes) are not necessarily influencers but removing them may be vital for the gradual increase of the average shortest path (but not for disconnecting a network) among the users in the hateful followers' networks and therefore, reduce of the hateful connectivity and exposure. However, due to the fluctuation of the increase and decrease of the average shortest path metric among the majority of the removal strategies for the followers' and retweets' networks, the author suggests that the average shortest path metric is an unsatisfactory measure for hateful networks' connectivity and flow when networks become significantly disconnected. This is congruent with the findings of other studies [54, 53].

Chapter 6 also attempted to provide an answer to the question:

RQ5: According to the structural characteristics of networks, would a combination of two (hybrid) node removal strategies be more effective at decreasing the propagation of hateful content - compared to applying a single node removal strategy?

A simulation of seven hybrid node-removal strategies demonstrated that the single node-removal strategies were more effective at reducing the largest component (GC) of hateful followers and retweets networks. This emphasises that hateful networks are effectively fragmented by removing the highest degree users (influential and super consumer/retweeters), more so than if any other roles in a network are removed. This means that removing nodes with the highest connection is the most efficient way to deconstruct such networks and combining other node roles may reduce this effect. This is in line with Bellingeri *et. al.*[40] who demonstrate that applying a combined attack strategy was the most efficient strategy to decrease the largest component size (GC) when deleting one or two important hubs. Nevertheless, the combined strategy was less efficient than the single strategy to deconstruct a network. However, the efficiency of node-removal strategies depends on the topology of the network [40], meaning that further investigation is required on different hateful networks. Chapter 6 also introduced further investigation around the disruption of hateful networks by applying all the previous node-removal strategies to bipartite hateful networks, attempting to provide an answer to the question:

RQ6: Would applying the node removal strategies to a bipartite version of hateful

networks improve the node removal strategies in terms of detecting the most important users?

To answer this question, the generic hateful followers/retweets networks were transformed into two-mode networks (bipartite networks). The experiments indicated that there was no advantage from applying 13 (single, hybrid and random) node-removal strategies to the bipartite networks compared to the results obtained by applying the single node-removal strategy to the generic networks. Perhaps this is because of some differences in these networks' characteristics. First, the overlap between cliques in the top nodes set is significant in the bipartite network, meaning that if two cliques have one node in common, then they certainly have many nodes in common[191]. This may be interpreted as indicating that it is hard to disconnect two cliques because they are connected to each other by many nodes. Moreover, the projection of the bipartite network (converting the two-mode into one-mode network to compare the GC of the bipartite network with the generic networks) may have resulted in many highly connected nodes [353], resulting in high clustering coefficients (high clustering coefficients in a projection may be seen as a consequence of the underlying bipartite projection rather than a specific property of the network [190]) and likewise, the projection may lead to very dense networks compared to the generic version, even if the bipartite version is not dense [190, 246]; these observations may confuse the results. Therefore, when performing social network analysis, one must be aware of the characteristics of the projected network to avoid the confusion of information obtained by the projected bipartite networks when compared to the generic version of that network. The experimental results led to the following contribution:

C4: *To the best of the author's knowledge, this is the first study to develop strategies that identify nodes within hateful networks (user accounts) whose removal is empirically shown to reduce the connectivity (largest component, density and average shortest path) in both follower and retweet networks. Thirteen node-removal strategies, in-*

cluding a random-based strategy based on network connectivity, were tested on three network metrics: giant component size, density and the average shortest path. These strategies were applied to generic networks and bipartite networks. The experiments carried out for this study demonstrate that the best node-removal strategy is the degree-based strategy (single node removal) which has the highest impact on reducing the size of the largest component of the generic hateful followers' and retweets' networks. The rigour of these findings is demonstrated for two hateful followers' networks and three hateful retweets' networks.

This study was based on the fact that no study yet exists that examines node removal strategies for multiple hateful Twitter networks, in terms of hate speech exposure (followers of networks) or hate speech propagation (retweeters of networks). Table 2.7, Chapter 2 demonstrated that the node removal strategies had been applied only on a political Twitter network while the hateful networks that have been disrupted are web forum networks (terrorist and criminal). These networks are topologically different from the hateful networks on Twitter. In addition, the table indicated that limited node removal strategies have been applied to hateful networks on the web while we applied a wide range of node-removal strategies (13 strategies).

7.2.2 The sufficiency and representativeness of the datasets:

The experimental results in Chapter 4 produced evidence to suggest that othering language can be a valuable feature for identifying cyberhate. However, would this finding generalise to a new dataset produced by, for example, a future anti-religious event? An experiment has been implemented (in Chapter 5) that showed some evidence that our models have had similar behaviour on unseen anti-religious datasets (Anti-Muslim 1/2). But in future, we have to accept the langauge use and therefore distribution of future cases of anti-religious content may vary. Training on new datasets in future might produce more generalisable models, but there is currently not enough consistency to determine which datasets or what properties of a dataset lead to more generalisable models [350]. Moreover, choosing the training dataset is as important as the choice of model. This because there is an assumption that the training data represent the distribution of future cases. It has been reported by [302, 172] that training the proposed model on Davidson's dataset (used for the dataset used for building the 'othering' feature set and training the classifier) leads to better generalisation for dissimilar datasets because of the large proportion of abusive posts (including those that are hateful and offensive). This is encouraging, yet more synthesis across different studies is needed for the recent dataset. For studies in Chapters 5 and 6, we tried to generalise the results by applying the experiments to multiple datasets. The experiments and the results referenced initial ideas about how hateful networks behave and the best solution for disrupting these networks. However, these experiments were conducted using a small sample of datasets. We should keep in mind that these datasets only cover current representative hateful networks, but one can expect that the diversity of network properties increases beyond this small sample.

7.3 Thesis Implications

Currently, there is neither a global definition of hate speech or online hate crimes, nor is there agreement on how to treat hateful online content [123]. As a result, there is much variation in how online social platforms police the problem, including an overall reluctance to remove hateful content [20]. Global understanding of (online) hate could therefore guide online social platforms on what content is considered hate and thus needs to be removed. Therefore, Chapter 2 is advantageous for the researcher and policymakers as it introduced extensive literature regarding the definition of cyberhate from different perspectives.

Concerning online hate classification, much research has been developed for detecting cyberhate using machine learning tools; however, what is needed is an improvement in the machine understanding of ’indirect hate’ rather than only detecting the occurrence of ’direct hate’. The study described in Chapter 4 sheds light on detecting hate speech using features that may not appear to be hateful out of context, using the concept of ’othering’ [62] which is a part of psychological theory.

The findings may widen researchers’ horizons to combine other psychological theories and state-of-the-art models, e.g. semantic learning, to improve machine learning to be closer to human judgement. This combination of machine learning automation and social theories may also be considered promising in the wider Artificial Intelligence (AI) field - developing feature sets integrating a range of psychological theories, e.g. Social Identity theory, Belonging theory, etc. Indeed, no research guarantees 100 percent accuracy in hate detection; this is because humans themselves might disagree on what is considered hate. Thus, developing cyberhate detection should continue and consider all hate stereotypes as features for classifying cyberhate. Direct hate provokes a reaction from online social platforms, e.g. account suspension. The concern remains that indirect hate that feeds people with hateful ideologies and discriminative ideas may still be left unmanaged.

In addition, understanding online connectivity in terms of exposure to and spreading hate is as important as detecting hateful content itself. On the Twitter platform, there is a lack of studies that help decision-makers respond to the spread of dangerous online content. The findings of Chapter 5 focused on identifying hateful networks’ connectivity and, therefore, may enable decision-makers to understand the risk of being exposed to online hate. Characterising hateful networks’ connectivity is a fundamental study that may motivate researchers to think of suitable management approaches. For example, the results of all the chapters combined may be beneficial for detecting, classifying or deconstructing hateful networks. Chapter 6 encouraged us to consider disrupting their connectivity and recorded changes in network connectivity. It can be argued that disrupting cyberhate propagation can mitigate hate consequences [228].

As these experiments were performed on hateful Twitter networks for the first time, it has opened the door for decision-makers to strengthen their disruption methods and prevent individuals from being influenced by hate groups at an earlier stage of developing such views. For example, it could benefit governments (e.g. KSA and the UK) and a range of organisations (e.g. charities and ISPs) in developing and implementing counter-hate/prevention strategies. Besides, it could aid the government in the implementation of the Prevent duty. For example, in the UK, the Prevent Strategy includes responding to the ideological challenge of terrorism, preventing people from being drawn into terrorism and working with a wide range of sectors and institutions (e.g. education) where there are risks of radicalisation which need to be addressed[1] [145]. For instance, we showed how engaging with users exhibiting high out-degree (those retweeting hate) could potentially lead to a reduction in hateful network size.

However, these strategies do not take the entire aspect of preventing expressions of online hate into account, as they do not address or explain the risks of hateful content to online users. Nevertheless, the importance of prevention has been acknowledged, both academically and politically. For example, Moghaddam *et al.* [228] argued that prevention is a long-term solution to terrorism.

Future research into the prevention of online hate may consider the role of all these insights for mitigating direct/indirect hate propagation. It should take the relationship between hate expression behaviour (direct/indirect) and its network, e.g. does the indirect hate expressed through othering language lead to detecting an anti-religious online network?

[1] https://www.lbhf.gov.uk/crime/prevent-strategy-overview-and-contact-details

7.4 Thesis General Limitations

7.4.1 Datasets' scarcity

Automated hate speech detection and analysing propagation depends on access to annotated content that humans agree to be hateful. In the field of cyberhate detection using Machine Learning (ML), datasets are used to evaluate or compare ML methods' performance on a particular task. It is used by researchers to test how their new ideas perform against existing ones [71] and to objectively measure progress on a particular problem. The dataset is usually the only necessary consistent/constant aspect of a study. The point here is that appropriate dataset collection methods are essential.

In the hate speech detection field, researchers tend to start by collecting and annotating new messages, and often these datasets remain unshared [118]. In the majority of papers focusing on 'algorithms for hate speech' new different data are collected and annotated.

Figure 7.1, by Fortuna *et al.* [118], showed that only in a few studies are data made available for other researchers (label 'own, available'), and only in one case is an already published dataset used ('published dataset').

The limited number of datasets that are publicly shared is a relevant aspect in the area of hate speech classification. This makes it difficult to compare results from different studies. This is largely due to the fact that while social media platforms facilitate discussions, including those about hate speech, many have strict data usage and distribution policies. This results in a relatively small number of datasets being available to the public to study.

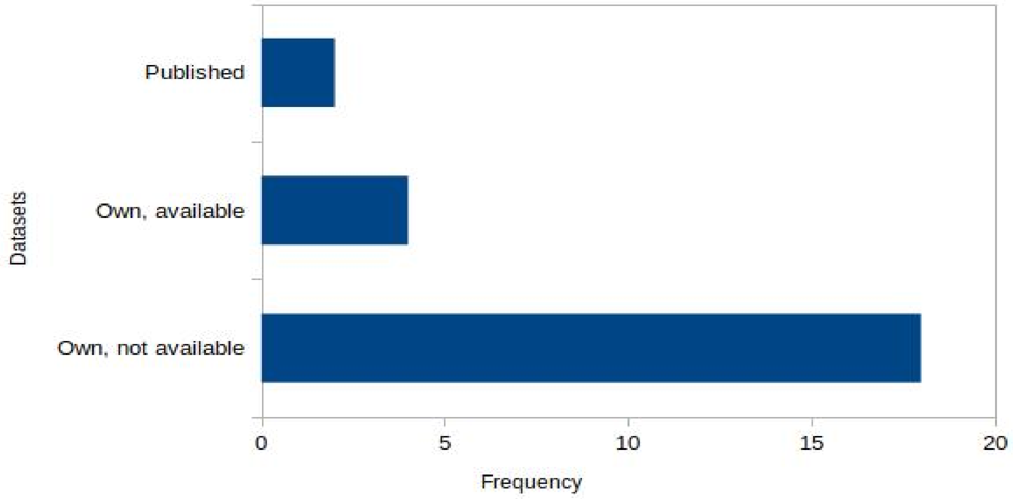


Figure 7.1: Dataset availability in the documents with algorithms about hate speech .

7.4.2 Text Annotation

In order to perform text classification experiments on hate speech detection, having access to labelled corpora is essential. Since there is no commonly accepted benchmark corpus for the task, authors usually collect and label their own data.

The reliability of human annotations is crucial, both to ensure that the algorithm can accurately learn the characteristics of hate speech, and as an upper bound on the expected performance [323]. A study performed by Ross *et al.* [276] on hateful content clarified that the agreement of the annotators was very low because they revealed that there is considerable ambiguity in existing definitions. A given statement may be considered hate speech or not depending on people's cultural background and personal sensibilities. This clarifies the problem of annotation when supervised learning is adopted. Nobata *et al.* [243] compared crowd-sourced annotations achieved using AMT with annotations created by expert annotators and found large differences in terms of agreement rate.

In addition to reliability problems, human annotation is costly when the research aims to examine big datasets. Furthermore, sometimes people, including those involved in our research, tend to annotate a number of samples (e.g. 2000) but the annotation results become imbalanced (e.g. 1800 for the benign and 200 for the hateful). Also, the status of hate speech is variable, which means that hateful instances may be considered non-hateful later on.

Collecting data from Twitter is affected by several factors. For example, several studies collected data by performing an initial manual search of common slurs and terms used pertaining to religious, sexual, gender, and ethnic minorities for the crowd-sourcing process. In the results, the collected data might contain tweets that were annotated differently.

To solve this, several studies (e.g.Waseem *et al.* [323]) provided a list of criteria founded in critical race theory, and used them to annotate a publicly available corpus. While this increases the proportion of hateful posts for resulting datasets, it focuses the resulting dataset on specific topics and certain sub types of hate speech (e.g. hate speech targeting Muslims). Furthermore, human annotation suffers from the problem of disagreement even though the criteria can be specified. This is because each person seems to intuitively sense what hate speech is, but rarely are two peoples' understandings the same [276]. For example, the tweet text 'Allah Akbar' would be annotated differently by different annotators from different cultures. 'Allah Akbar' is a common Islamic Arabic expression, used in various contexts by Muslims, usually informal pray 'Salah'. However, this expression existed in a lexicon used for detecting hate speech by [96]. This causes the issue of small and imbalanced samples of hateful text and non-hateful text within a dataset.

Another problem is the size of the annotated dataset. In general, the size of a collected corpus varies depending on specific works on hate speech detection, ranging from around 100 labelled comments [103, 106], to several thousand comments used in other work, such as [323].

7.4.3 Twitter Data

There are often specific challenges with using social media data in academic research, and in particular Twitter data. Below is a list of some of the challenges that have been faced when using Twitter as a data source in this research.

- Sharing of tweets is prohibited under Twitter's API Terms of Service, though researchers can share tweet identification numbers, associated with each tweet, which can be used by other researchers to obtain Twitter datasets. See Twitter's API Terms of Service [4]. However, Curini *et al.* [93] showed two problems related to using tweet identification numbers. First, recovering tweets by their identification number needs programming skill that not all researchers are familiar with. Second, if a user deleted the tweet or the Twitter account, the tweet will no longer be available and recoverable.

- There are ethical issues if a researcher decides to reproduce tweets in an academic publication, especially tweets related to sensitive topics. Researchers should gain informed consent from the user posting the tweet, which is difficult to obtain simply due to the volume of tweets retrieved. See Williams *et al.* [333] and Beninger's *et al.*[44] reports on users' views of research using social media. This limits this type of research in terms of more detailed textual analysis of hateful tweets.

- With regard to Chapters 5 and 6, it was noticed that the annotated hateful tweets produced a number of suspended accounts (which violated Twitter's policy and were reported by other users). According to Twitter's suspension policy, users may not use the platform for the purposes of spamming, yet there are different reasons for account suspension that include: 1- account security at risk, e.g. an account has been hacked or compromised; 2- abusive tweets or behaviour, e.g. sending threats to others or impersonating other accounts [316]. Twitter provides an option for people to report an account that results in violence and they were asked to provide several tweets from that account to better understand the context. In the collection under study here, a number of suspended accounts were discarded because their profiles could not be collected, e.g. followers collection, which was a basic step for building the hateful followers' networks.

7.5 Future Work

7.5.1 Develop Larger Datasets

In future, larger datasets could be developed on which to test the 'othering' classifiers and could be used to study rising and falling cyberhate levels on a range of online social media platforms, with the intention of collecting the 'othering' narratives to better understand the topics and touchpoints being discussed in this context at an aggregate level during times of civil unrest or following trigger events. Also, for the network characteristics, more extensive hateful datasets could be constructed and compared to the baseline study used in this work, to establish consistency of findings across a larger range of hateful networks and to study differences over time. Where possible, this may also include other online social networks with open APIs and it may include a comparison with different 'risky' networks (e.g. harassment and cyberbullying). Furthermore, the cost of human annotation limits the size of the annotated posts.

In the future, we may see the use of machine classifiers for detecting cyberhate applied to larger datasets, with sub-sampling to validate performance. Current approaches are continually improving in performance up to 96% accuracy [25, 202, 203].

7.5.2 Improving Feature Sets:

The feature set is an individual measurable property or characteristic of a phenomenon being observed, e.g. in Chapter 4 *'the othering feature set'*. Choosing informative, discriminating and independent features is a crucial step for effective algorithms in classification. Thoughts for future work relate to network node characteristics as an additional feature that could be combined with the ’othering feature’ for detecting hateful content and also hateful users. For example, a future research question could be - do people who post hateful content have similar network metrics? If the answer to that question is ’yes’, this means that the hateful users’ position in a hateful network could be a feature utilised for training a classifier to detect hateful content or users. Previous studies used network characteristics for detecting hateful users [75, 273, 310] but no study has yet exploited these characteristics as a feature in combination with the feature in this work; see Chapter 4, for the detection of hateful content itself.

7.5.3 Implicit content and their networks

In the future, the propagation of hateful networks should be compared and contrasted based on two sorts of users: (i) users who have posted implicit hate (e.g. ’othering’ content) and (ii) users who have posted explicit hate (e.g. dirty ni***a). These two networks are sub-networks of the main hateful network. The importance of this investigation would be to measure the level of similarities among these networks and the differences based on the following questions: (i) which sort of users’ network is denser, that of users who propagate implicit hate or explicit hate? (ii) which sort of network has a smaller average shortest path? (this will clarify which sort of hateful speech propagates faster) and (iii) which sub-network has higher reciprocity behaviour? Do users who publish the implicit hate retweet each other, or are users who publish the explicit hate more cooperative in terms of retweeting? In addition, it is possible to measure the correlation between the implicit sub-network and the explicit sub-network. If the correlation is low, this means that we are facing two groups of haters: haters who prefer to publish only implicit hate and users who prefer to publish only explicit hate.

7.5.4 Disrupting Multipartite Networks

A multipartite network is a network whose nodes are or can be partitioned into different independent sets, forming a partition such that no two nodes belonging to the same subset are adjacent. Despite the fact that bipartite networks have not previously offered promising results in terms of identifying the important node, in the future, we will apply the node removal strategies to our hateful networks transformed into multipartite networks. In addition, we will apply edge-removal strategies and compare and contrast the results with the node-removal strategies. A matrix of the networks' mode (column) and the node/edge-removal strategies (row) will be helpful for the literature as a source that measures the performance of the node/edge removal strategies in terms of disconnecting hateful networks.

7.5.5 Hidden networks: The propagation of hash-tagged tweets

Future work could investigate the inclusion of hashtags in tweets in the context of studying hate speech propagation on the web, in particular, a comparison of hate propagation with hash-tagged tweets versus non-hash-tagged tweets. This is motivated by the observation that the inclusion of hashtags can increase the chance of tweet propagation [92, 194]. Future research could ask - does the inclusion of the hashtags contribute positively to the propagation of hateful content? This is important information to clarify because the use of hashtags in tweets might well build a 'hidden' network of hateful users that targets their ideologies to specific groups and topics.

7.6 Summary

In this thesis, it has been demonstrated that exploiting the use of pronouns in discriminating content as a feature improves the training results of the state-of-the-art machine classifiers and also detecting hate in unseen datasets. Also, extensive analysis was conducted to reveal the similarities between hateful networks and strategies for the disruption of hate propagation.

Appendices

Appendix

Appendix A: Recent Literature Review

Since hate speech and the abusive language have recently become subjects of general concern, detecting hate speech has grown to be a major topic by the community of natural language processing (NLP), as demonstrated by the creation of datasets in a variety of languages [250]. Therefore, there are a considerable numbers of hate speech detection and classification have been published after our contribution, the period from 2019 to 2021. The table below is a continue of Table 2.5 in Chapter 2.

Table A1: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Study	**Hate Type**	**Year**	**classifier**	**Features**	**Accuracy**	**Precision**	**Recall**	**F-score**	**AUC**
Fatahillah ***et al.***[113]	Hate speech	2019	Multinomial Logistic Regression	TF-IDF	0.87	0.80	0.82	-	-

Table A1 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Van *et al.*[314]		2019	Bi-GRU-CNN and BiGRU-LSTM-CNN	TextCNN	-	-	-	0.70	-
Pamungkas*et al.*[254]		2019	LSVC, LSTM and HurtLex	Word embedding	-	0.60	0.79	0.68	-
Rodriguez *et al.*[274]	Hate speech	2019	K-means	VADER and JAMMIN, TF-IDF	0.74	-	-	-	-
Liu*et al.*[202]	Hate speech	2019	Fuzzy classifier	Embedding features	-	-	-	0.93	-
Briliani *et al.*[59]	hate speech	2020	K-Nearest Neighbour	TF-IDF	0.97	0.93	0.94	0.93	-

Table A1 Continued: Summary of the studies that related to hateful text classification showing the hate type, year, classifiers used, metrics accuracy, precision, recall, f-measure and AUC

Oriola*et al.*[247]	Offensive and hate speech	2020	Optimized Gradient Boosting	n-gram	TF-IDF	0.93	-	-	-
Alsafari ***et al.***[28]		2020	CNN and mBert	CNN	-	0.78	0.76	0.81	-
Chatterjee ***et al.***[74]	Cyberbullies	2020	Fuzzy classifier	Bandpower (Bp)	0.90	-	-	-	-
Ayo ***et al.***[35]	Hate speech	2021	Fussy classifier	TF-IDF	-	-	-	0.92	0.96
Miok*et al.*[221]	Offesnsive language	2021	BERT and Monte Carlo dropout (MCD)	BERT	0.91	-	-	0.90	-

In 2019, Fatahillah et al. [113] collected the data from Twitter and employed Case Folding, Tokenizing, Filtering, and Stemming methods in preprocessing phase. After Pre-processing, the TF-IDF technique is used for vectorization. After Feature engineering, the Logistic regression algorithm has been applied, and they have found 87% of accuracy. Axel Rodriguez et al. [274] proposed an approach to detect hate speech content using sentiment analysis on Facebook. They used Graph API to extract the post and comments from Facebook. To remove the unrelated texts VADER and JAMMIN were used. In preprocessing phase, they filtered out all unnecessary stopwords or symbols. Preprocessed documents converted into the vector using TFIDF. The resulting matrix is passed to the k-means clustering algorithm as an input matrix. The most negative articles and responses were collected using sentiment and emotion analysis. Tin Van Huynh et al. [314] proposed an approach to detect hate speech using Bi-GRU-CNN-LSTM Model. In this paper, they collected data from Twitter and categorized their data into three labels (OFFENSIVE, HATE, and CLEAN). After cleaning the data, they implemented three neural network models such as BiGRU-LSTM-CNN, Bi-GRU-CNN, and TextCNN to identify hate speech. They achieved a 70.57% of F1 score as a result.

Viviana Patti et al. [254] proposed a Hybrid based approach to detect hate speech [31]. In this paper, they employed two models. In their first model, they implemented a linear support vector classifier (LSVC), and in the second model, they employed a long short-term memory (LSTM) neural model with word embedding. They concatenated 17 categories, such as HurtLex, with two types, namely LSVC and LSTM. Joint learning with a multilingual word embedding model, including HurtLex, performed best with 68.7% of F1-score.

Nanduri et. al.[235] created a method for classifying the multi-class instances using the fuzzy methodologies. They attempted for designing an updated fuzzy that includes two stages of training for the classification of cyberhate conversation into 4 forms, race, disability, sexual orientation, and religion. By performing several experiments on the present process the experimental and theoretical estimation has validated the characteristics through ML approaches such as the words that combined for future methods Bow (bag-of-words) and extraction, which also differentiate cyberhate and normal conversation very appropriately.

Another study by Chatterjee et. al.[74] suggests a method for identifying and classifying cyberbullying acts as harassment, flaming, terrorism, and racism. The author uses a fuzzy classification rule; therefore, the results are inferior in terms of accuracy (around 40%), but using a set of rules, improved the classifier efficiency by up to 90%.

Liu et. al. [202] introduced a novel formulation of the hate speech type identification problem in the setting of multi-task learning through our proposed fuzzy ensemble approach. In this setting, single-labelled data can be used for semi-supervised multi-label learning and two new metrics (detection rate and irrelevance rate) are thus proposed to measure more effectively the performance for this kind of learning tasks. They reported an experimental study on identification of four types of hate speech, namely: religion, race, disability and sexual orientation. The experimental results show that our proposed fuzzy ensemble approach outperforms other popular probabilistic approaches, with an overall detection rate of 0.93.

In 2020, Oluwafemi Oriola et al. [247] proposed an approach to detect offensive speech on tweeter. The author collected the data set using Twitter API and annotated those data set into two sections, free speech âFSâ and hate speech âHT.â In preprocessing phase, they removed special characters, emojis, punctuations, symbols, hashtags, stopwords to clean the data. In the feature engineering phase, they employed the TF-IDF technique to transform the text into feature vectors. After applying an optimized support vector machine with n-gram, they have found 89.4% of accuracy.

Annisa Briliani et al. [59] proposed an approach to identify hate speech on Instagram using the k-nearest neighbour classifier. They collected the data set using Instagram API from Instagram and annotated those data set manually. They divided the dataset into 2 labels, namely zero and one. In preprocessing phase, they cleaned the data and employed the TF-IDF technique in the feature engineering phase. After then, they applied the k-nearest neighbor algorithm and found 98.13% of accuracy.

Alsafari et al. [28] proposed a Hate speech detection model for Arabic social media. In this paper, they collected the data set using Twitter search API, and the data set is categorized into four classes (Religious, Nationality, Gender, and Ethnicity). They cleaned the data set in preprocessing phase by removing unnecessary words such as URLs, punctuations, symbols, tags, and stopwords. They implemented a three-class classification with CNN and Bert to achieve 75.51% of the F1-score.frequent validation or on demand validation - both can generate considerable, often unnecessary, network traffic and the latter reduces much of the latency gains offered by caching. The viable alternative in such circumstances is resource-driven invalidation where the server invokes a callback on the cache to inform it whenever an update has

In 2021, Ayo et. al.[35] suggested a probabilistic clustering model for hate speech classification in twitter which tackle the problem of emotions overlap between positive or negative class. Features representation was done with Term Frequency- Inverse Document Frequency (TF-IDF) model and enhanced with topics inferred by a Bayes classifier. A rule-based clustering method was used to automatically classify real-time tweets into the correct topic clusters. Fuzzy logic was then used for hate speech classification using semantic fuzzy rules and a score computation module. From the evaluation results, it was observed that the developed model performed better in hate speech detection with F1-sore of 0.9256 using a 5-fold cross validation.

Calderon [67] Used unsupervised topic modeling to characterize hate speech against immigrants on Twitter in Spain around the appearance of the far-right party Vox. They concluded that the hate speech against immigrants produced around Vox, and not necessarily by Vox, followed the general patterns of this type of speech detected in previous works, including Islamophobia, offensive language more often than violent language, and the refusal to offer public assistance to these collectives.

Miok et. al. [221] proposed a Bayesian method using Monte Carlo dropout features within the attention layers of the transformer models to provide well-calibrated reliability estimates. Their experiments show that Monte Carlo dropout provides a viable mechanism for reliability estimation in transformer networks. Used within the BERT model, it offers state-of-the-art classification with performance of 0.91 accuracy and 0.90 F-score and can detect less trusted predictions.

Appendix B:Further analysis for hate speech definitions

Table A2: Analysing the hate speech definition according to type of the hate expression

Definition reference	**Attack**	**incite**	**harm**	**diminsh**	**discri-minate**	**Justify hatered**	**Humour**	**Thr-eaten**	**Hostility**	**Demean**	**Spread hate**
Bansal et al.[38]	✓	✓	✓	✓	✓	✓	✓	X	X	X	X
ILGA[127]	X	✓	X	X	✓	✓	X	X	✓	X	✓
Code of Conduct between EU and companies[10]	X	✓	X	X	X	X	X	X	X	X	X
Tarasova et al.[306]	X	X	X	X	X	X	X	X	✓	X	X
Nobata et al.[243]	✓	X	X	X	X	X	X	X	X	✓	X
Twitter[4]	✓	✓	X	X	X	X	X	✓	X	X	X
YouTube[3]	X	✓	X	X	X	X	X	X	X	X	X
Mondal et al.[229]	✓	X	X	X	X	X	✓	X	X	X	X
Fortuna et al.[118]	✓	✓	X	✓	X	X	✓	X	X	X	X
Encyclopedia of the American Constitution[244]	✓	X	X	X	X	X	X	X	X	X	X
De et al.[99]	✓	X	X	X	X	X	X	X	X	X	X
Thesis's definition	✓	✓	✓	X	✓	✓	X	✓	X	X	✓

Appendix C: Examples of hate speech features

Table A3: The most similar words to the othering terms based on four hate speech types. .

	Trained Data Set	Test Data Sets			
words	Othering Feature Set	Religion	Race	Disability	Sexual-oreintation
us	ni**as, we, Arab, those, iran, group, Jews, Muslim, send, headiest,	these, nig**s, they, African, Arab, them, you, blacklist, violence	UK, state, canada, they, me, Germany, Iraq, them, out of	wife, gurentee, bought, final, Poland, stew, we, he	did, make, we, you, get, athelet, nig**a, brave
them	these, Jew, those, pakistanian someon hang, Muslims, rednecks, country, p**s, niggu	these, you, p**s, send, those, kick, fuck, muslims, Arab	their, us, these, those, many, themselves, stop, country, ni**aas	short, bald, though, smash, yourself, never, powel, retain	they, cunt, fu**, Fu***ng, their, these, idiot, chrestian
Muslim	us, nonmuslim, iraq, lable, Jews, racism high, arab, yellow, doe, country	crime, Jews, Pakistani, blacklash, socity, christian, white, nonmuslim	Islamic, Arab, Sunni, Jews, us, Iraq, christian, hindu minority, religious	Not exist	Not exist
out	gotta, outta, Islam, America, chop, attack, mosque, send home, manchest, blacks, these	nazzi, carri, move, lot, matter,them, asian, gang, send, black	down, back, their, them, these, wrong, fight, blacks, US, building	juturna, they, primlyst, reason, miss, mild, fatty,you, smash	announce, these, ni**o, active, black, community, move
burn	church, non-whites, chop, mosques, themfu**ing, shoot, chop, ni**as, commit, nigeria	church, story, mosques, Saudi, action, distance, fu**ing, nig**ro	kill, destroy, fu**, explosion, shoot, strick, obama, nigeria voting	Not exist	Not exist

Examples of the whole typed dependency sentences from anti-religious dataset:

1- nig**isss bare peopl say one of the peopl that kill the man in woolwich live behind***

root(ROOT-0, Nig*****isss-1)

amod(people-3, bare-2)

nsubj(NiggaGenesisss-1, people-3)

nsubj(killed-10, people-3)

acl(people-3, saying-4)

dobj(saying-4, one-5)

case(people-8, of-6)

det(people-8, the-7)

nmod:of(one-5, people-8)

ref(people-3, that-9)

acl:relcl(people-3, killed-10)

det(man-12, the-11)

dobj(killed-10, man-12)

case(lives-15, in-13)

amod(lives-15, woolwich-14)

nmod:in(killed-10, lives-15)

case(u-17, behind-16)

nmod:behind(killed-10, u-17)

2- ***told me Gary want the per that shot of them dirti black cunt that kill that lad in woo***

compound(Gary-4, brentthomas-1)

compound(Gary-4, BigJoeLew-2)

compound(Gary-4, garryt-3)

nsubj(told-5, Gary-4)

root(ROOT-0, told-5)

nsubj(cunts-17, me-6)

advmod(cunts-17, u-7)

cop(cunts-17, was-8)

det(cunts-17, the-9)

amod(cunts-17, per-10)

det(shot-12, that-11)

dep(per-10, shot-12)

case(them-14, of-13)

nmod:of(shot-12, them-14)

amod(cunts-17, dirty-15)

amod(cunts-17, black-16)

ccomp(told-5, cunts-17)

nsubj(killed-19, cunts-17)

ref(cunts-17, that-18) acl:relcl(cunts-17, killed-19)

det(lad-21, that-20) dobj(killed-19, lad-21)

case(woo-23, in-22) nmod:in(killed-19, woo-23)

3- fuckfem I wa told that woolwich thing wa a pakistani too loool go back to your countri pal or jump off a cliff either is g

root(ROOT-0, fuckfems-1)

dobj(fuckfems-1, I-2)

compound(Told-4, Was-3)

nsubj(fuckfems-1, Told-4)

mark(LOOOL-12, That-5)

compound(Thing-7, Woolwich-6)

nsubjpass(LOOOL-12, Thing-7)

auxpass(LOOOL-12, was-8)

det(pakistani-10, a-9)

nmod:npmod(too-11, pakistani-10)

advmod(LOOOL-12, too-11)

dep(fuckfems-1, LOOOL-12)

ccomp(LOOOL-12, go-13)

compound:prt(go-13, back-14)

case(pal-18, to-15)

nmod:poss(pal-18, your-16)

compound(pal-18, country-17)

nmod:to(go-13, pal-18)

cc(go-13, or-19)

ccomp(LOOOL-12, jump-20)

conj:or(go-13, jump-20)

compound:prt(jump-20, off-21)

det(cliff-23, a-22)

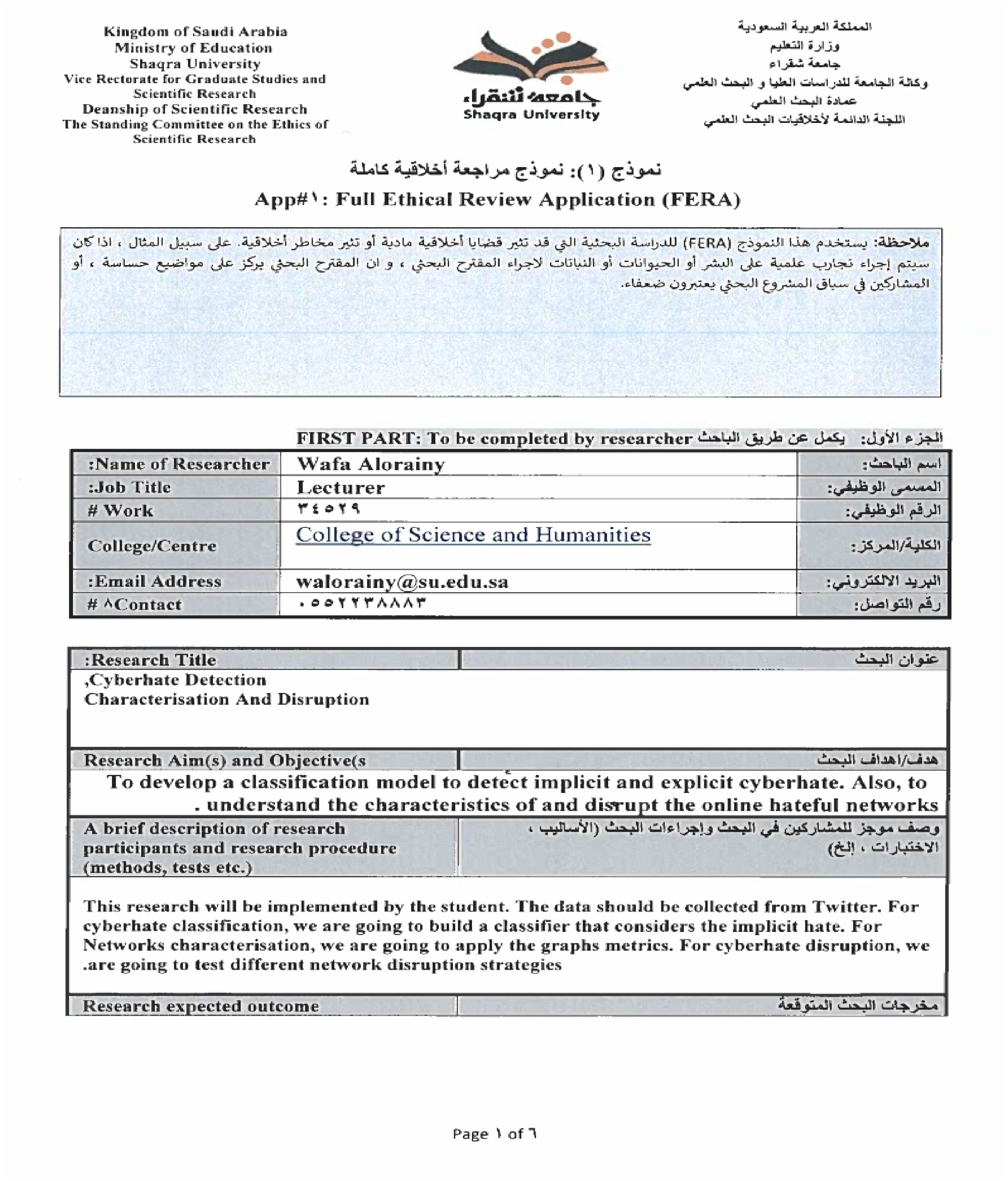
dobj(jump-20, cliff-23)

dep(LOOOL-12, either-24)

cop(g-26, is-25)

conj(LOOOL-12, g-26)

Appendix D: Ethical Approval



Kingdom of Saudi Arabia
Ministry of Education
Shaqra University
Vice Rectorate for Graduate Studies and Scientific Research
Deanship of Scientific Research
The Standing Committee on the Ethics of Scientific Research

جامعة شقراء
Shaqra University

المملكة العربية السعودية
وزارة التعليم
جامعة شقراء
وكالة الجامعة للدراسات العليا و البحث العلمي
عمادة البحث العلمي
اللجنة الدائمة لأخلاقيات البحث العلمي

نموذج (١): نموذج مراجعة أخلاقية كاملة

App#١: Full Ethical Review Application (FERA)

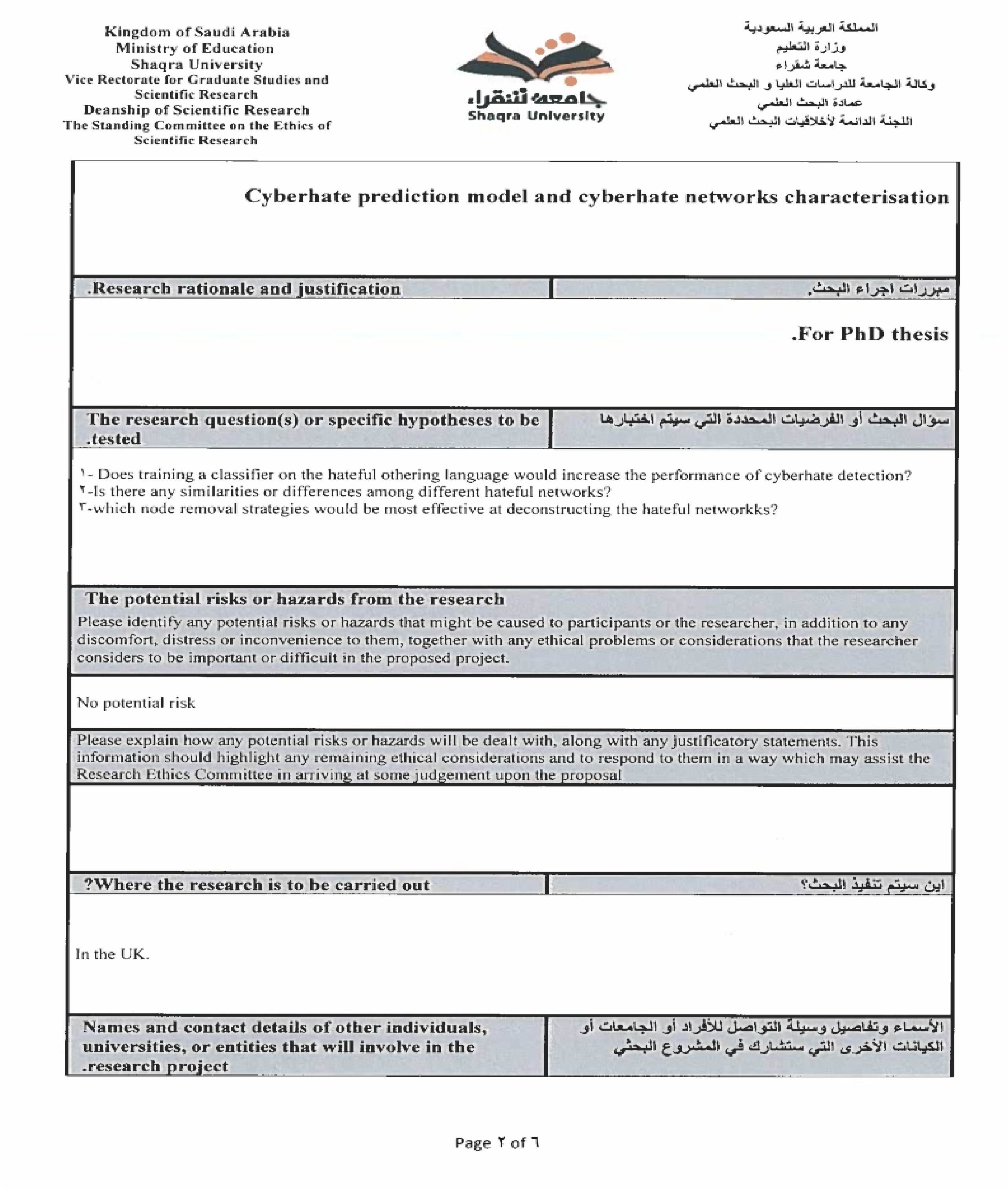
ملاحظة: يستخدم هذا النموذج (FERA) للدراسة البحثية التي قد تثير قضايا أخلاقية مادية أو تثير مخاطر أخلاقية. على سبيل المثال ، اذا كان سيتم إجراء تجارب علمية على البشر أو الحيوانات أو النباتات لاجراء المقترح البحثي ، و ان المقترح البحثي يركز على مواضيع حساسة ، أو المشاركين في سياق المشروع البحثي يعتبرون ضعفاء.

FIRST PART: To be completed by researcher يكمل عن طريق الباحث :الجزء الأول

:Name of Researcher	Wafa Alorainy	اسم الباحث:
:Job Title	Lecturer	المسمى الوظيفي:
# Work	٣٤٥٢٩	الرقم الوظيفي:
College/Centre	College of Science and Humanities	الكلية/المركز:
:Email Address	walorainy@su.edu.sa	البريد الالكتروني:
# ٨Contact	٠٥٥٢٢٣٨٨٨٣	رقم التواصل:

:Research Title	عنوان البحث
,Cyberhate Detection Characterisation And Disruption	
Research Aim(s) and Objective(s	هدف/اهداف البحث
To develop a classification model to detect implicit and explicit cyberhate. Also, to . understand the characteristics of and disrupt the online hateful networks	
A brief description of research participants and research procedure (methods, tests etc.)	وصف موجز للمشاركين في البحث وإجراءات البحث (الأساليب ، الاختبارات ، إلخ)
This research will be implemented by the student. The data should be collected from Twitter. For cyberhate classification, we are going to build a classifier that considers the implicit hate. For Networks characterisation, we are going to apply the graphs metrics. For cyberhate disruption, we .are going to test different network disruption strategies	
Research expected outcome	مخرجات البحث المتوقعة

Page ١ of ٦



Kingdom of Saudi Arabia
Ministry of Education
Shaqra University
Vice Rectorate for Graduate Studies and Scientific Research
Deanship of Scientific Research
The Standing Committee on the Ethics of Scientific Research

المملكة العربية السعودية
وزارة التعليم
جامعة شقراء
وكالة الجامعة للدراسات العليا و البحث العلمي
عمادة البحث العلمي
اللجنة الدائمة لأخلاقيات البحث العلمي

Cyberhate prediction model and cyberhate networks characterisation

.Research rationale and justification	مبررات اجراء البحث.

.For PhD thesis

The research question(s) or specific hypotheses to be .tested	سؤال البحث أو الفرضيات المحددة التي سيتم اختبارها

١- Does training a classifier on the hateful othering language would increase the performance of cyberhate detection?
٢-Is there any similarities or differences among different hateful networks?
٣-which node removal strategies would be most effective at deconstructing the hateful networkks?

The potential risks or hazards from the research

Please identify any potential risks or hazards that might be caused to participants or the researcher, in addition to any discomfort, distress or inconvenience to them, together with any ethical problems or considerations that the researcher considers to be important or difficult in the proposed project.

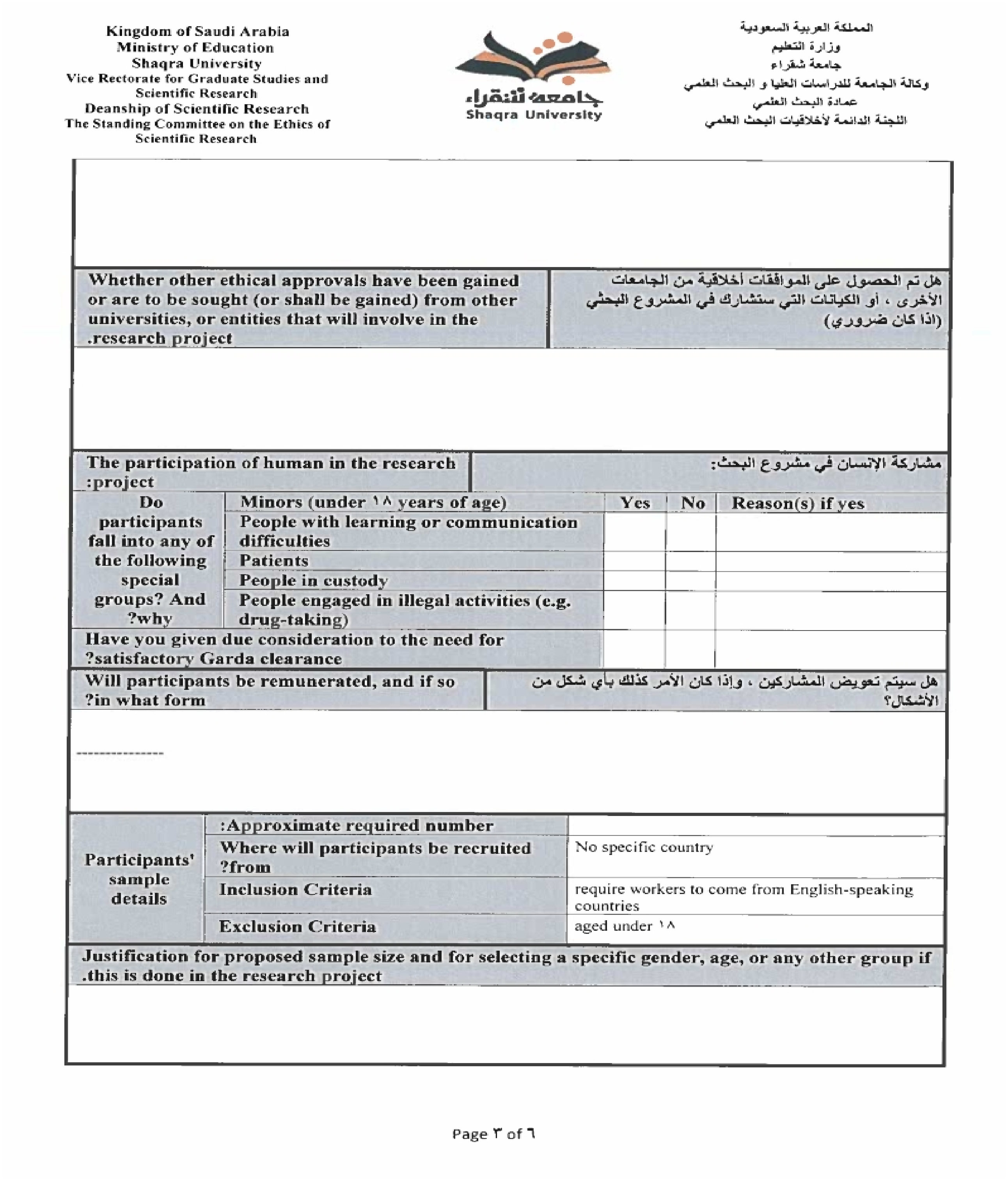
No potential risk

Please explain how any potential risks or hazards will be dealt with, along with any justificatory statements. This information should highlight any remaining ethical considerations and to respond to them in a way which may assist the Research Ethics Committee in arriving at some judgement upon the proposal

?Where the research is to be carried out	اين سيتم تنفيذ البحث؟

In the UK.

Names and contact details of other individuals, universities, or entities that will involve in the .research project	الأسماء وتفاصيل وسيلة التواصل للأفراد أو الجامعات أو الكيانات الأخرى التي ستشارك في المشروع البحثي



Kingdom of Saudi Arabia
Ministry of Education
Shaqra University
Vice Rectorate for Graduate Studies and Scientific Research
Deanship of Scientific Research
The Standing Committee on the Ethics of Scientific Research

جامعة شقراء
Shaqra University

المملكة العربية السعودية
وزارة التعليم
جامعة شقراء
وكالة الجامعة للدراسات العليا و البحث العلمي
عمادة البحث العلمي
اللجنة الدائمة لأخلاقيات البحث العلمي

Whether other ethical approvals have been gained or are to be sought (or shall be gained) from other universities, or entities that will involve in the .research project	هل تم الحصول على الموافقات أخلاقية من الجامعات الأخرى ، أو الكيانات التي ستشارك في المشروع البحثي (اذا كان ضروري)
The participation of human in the research :project	مشاركة الإنسان في مشروع البحث:

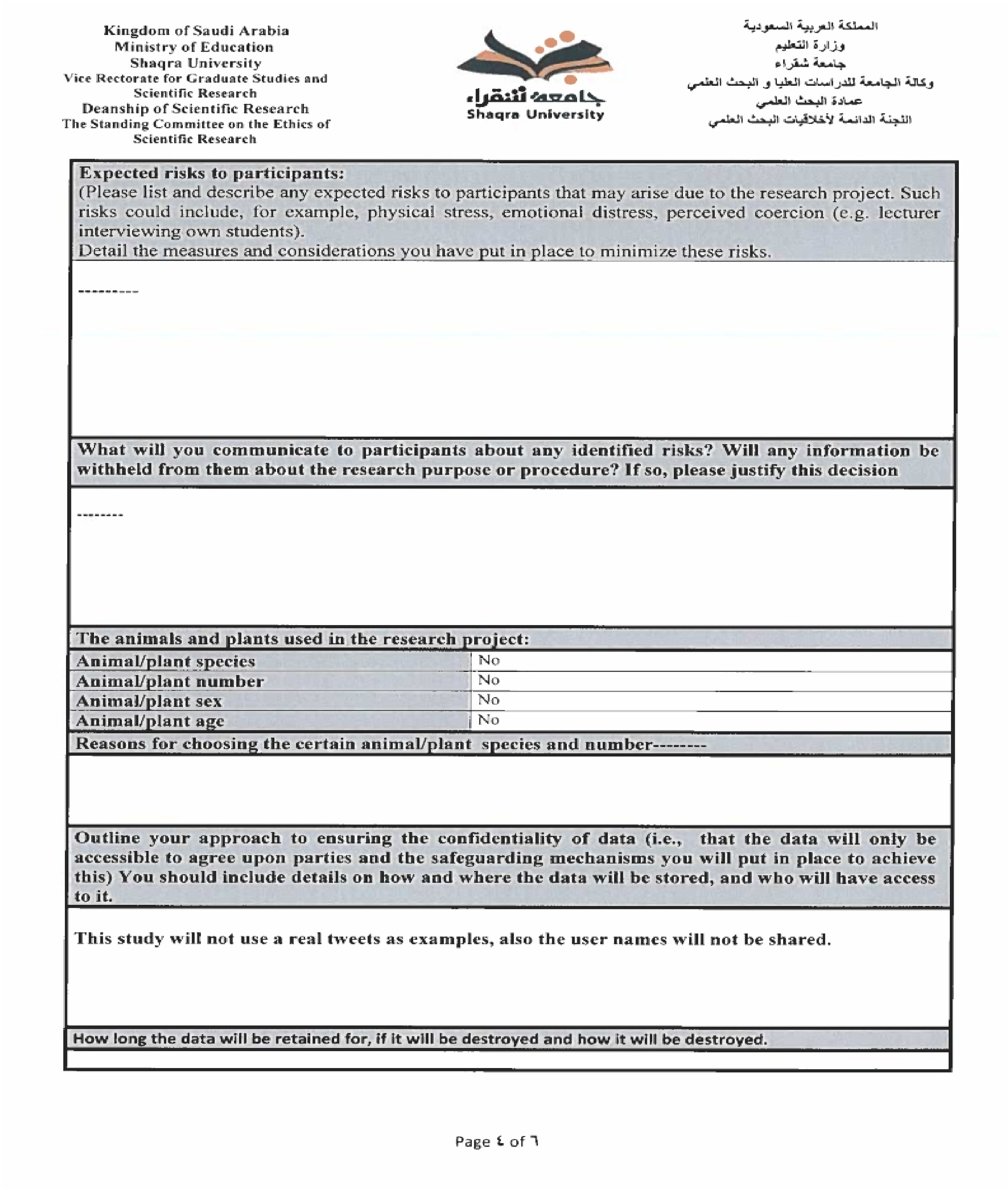
Do participants fall into any of the following special groups? And ?why	Minors (under ١٨ years of age)	Yes	No	Reason(s) if yes
	People with learning or communication difficulties			
	Patients			
	People in custody			
	People engaged in illegal activities (e.g. drug-taking)			
Have you given due consideration to the need for ?satisfactory Garda clearance				

Will participants be remunerated, and if so ?in what form	هل سيتم تعويض المشاركين ، وإذا كان الأمر كذلك بأي شكل من الأشكال؟

Participants' sample details	:Approximate required number	
	Where will participants be recruited ?from	No specific country
	Inclusion Criteria	require workers to come from English-speaking countries
	Exclusion Criteria	aged under ١٨

Justification for proposed sample size and for selecting a specific gender, age, or any other group if .this is done in the research project

Page ٣ of ٦



Kingdom of Saudi Arabia
Ministry of Education
Shaqra University
Vice Rectorate for Graduate Studies and Scientific Research
Deanship of Scientific Research
The Standing Committee on the Ethics of Scientific Research

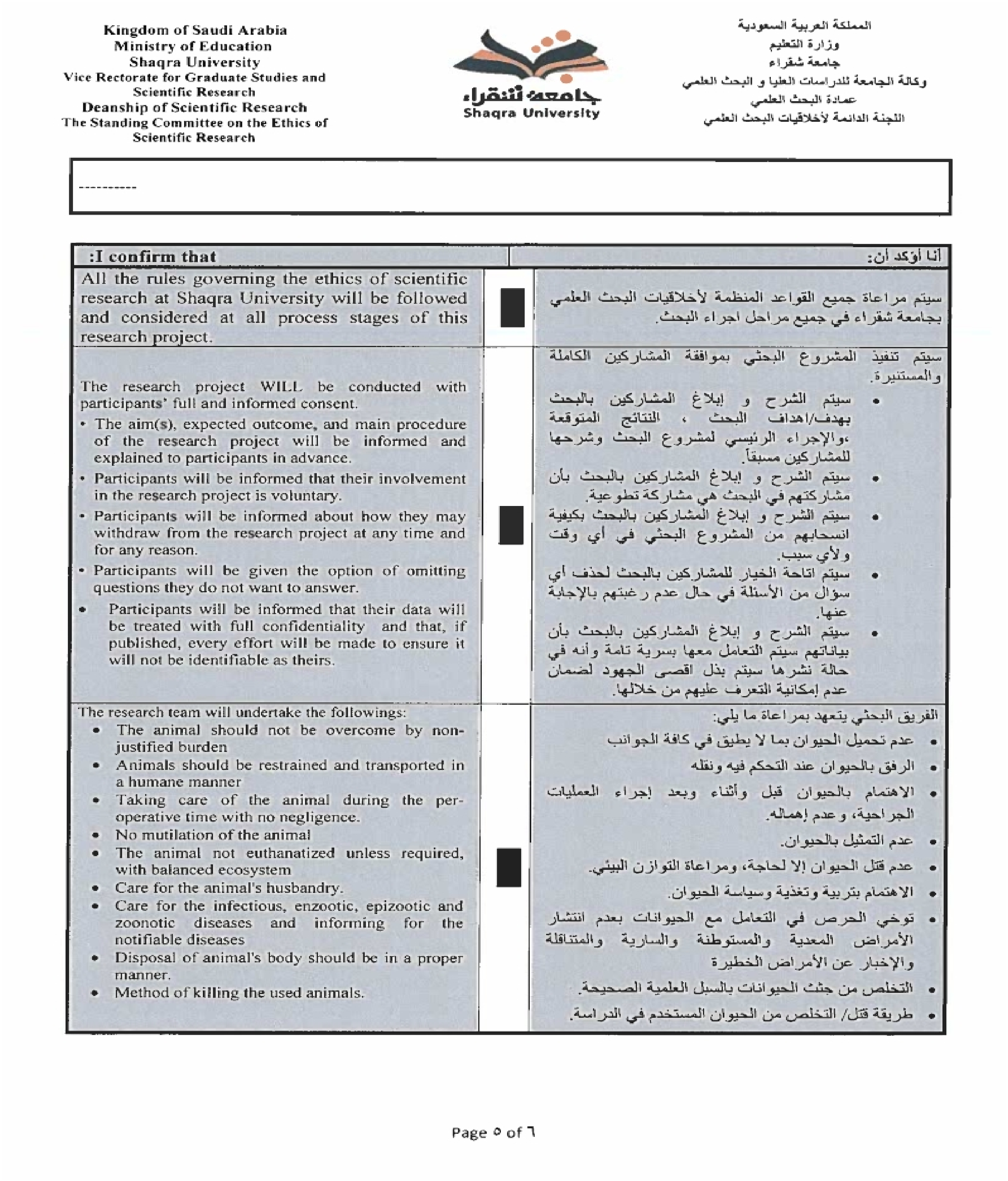
Shaqra University

المملكة العربية السعودية
وزارة التعليم
جامعة شقراء
وكالة الجامعة للدراسات العليا و البحث العلمي
عمادة البحث العلمي
اللجنة الدائمة لأخلاقيات البحث العلمي

Expected risks to participants: (Please list and describe any expected risks to participants that may arise due to the research project. Such risks could include, for example, physical stress, emotional distress, perceived coercion (e.g. lecturer interviewing own students). Detail the measures and considerations you have put in place to minimize these risks.

What will you communicate to participants about any identified risks? Will any information be withheld from them about the research purpose or procedure? If so, please justify this decision
----

The animals and plants used in the research project:	
Animal/plant species	No
Animal/plant number	No
Animal/plant sex	No
Animal/plant age	No
Reasons for choosing the certain animal/plant species and number--------	
Outline your approach to ensuring the confidentiality of data (i.e., that the data will only be accessible to agree upon parties and the safeguarding mechanisms you will put in place to achieve this) You should include details on how and where the data will be stored, and who will have access to it.	
This study will not use a real tweets as examples, also the user names will not be shared.	
How long the data will be retained for, if it will be destroyed and how it will be destroyed.	



Kingdom of Saudi Arabia
Ministry of Education
Shaqra University
Vice Rectorate for Graduate Studies and Scientific Research
Deanship of Scientific Research
The Standing Committee on the Ethics of Scientific Research

جامعة شقراء
Shaqra University

المملكة العربية السعودية
وزارة التعليم
جامعة شقراء
وكالة الجامعة للدراسات العليا و البحث العلمي
عمادة البحث العلمي
اللجنة الدائمة لأخلاقيات البحث العلمي

:I confirm that		أنا أؤكد أن:
All the rules governing the ethics of scientific research at Shaqra University will be followed and considered at all process stages of this research project.	■	سيتم مراعاة جميع القواعد المنظمة لأخلاقيات البحث العلمي بجامعة شقراء في جميع مراحل اجراء البحث.
The research project WILL be conducted with participants' full and informed consent. • The aim(s), expected outcome, and main procedure of the research project will be informed and explained to participants in advance. • Participants will be informed that their involvement in the research project is voluntary. • Participants will be informed about how they may withdraw from the research project at any time and for any reason. • Participants will be given the option of omitting questions they do not want to answer. • Participants will be informed that their data will be treated with full confidentiality and that, if published, every effort will be made to ensure it will not be identifiable as theirs.	■	سيتم تنفيذ المشروع البحثي بموافقة المشاركين الكاملة والمستنيرة. • سيتم الشرح و إبلاغ المشاركين بالبحث بهدف/اهداف البحث ، النتائج المتوقعة ،والإجراء الرئيسي لمشروع البحث وشرحها للمشاركين مسبقاً. • سيتم الشرح و إبلاغ المشاركين بالبحث بأن مشاركتهم في البحث هي مشاركة تطوعية. • سيتم الشرح و إبلاغ المشاركين بالبحث بكيفية انسحابهم من المشروع البحثي في أي وقت ولأي سبب. • سيتم اتاحة الخيار للمشاركين بالبحث لحذف اي سؤال من الأسئلة في حال عدم رغبتهم بالإجابة عنها. • سيتم الشرح و إبلاغ المشاركين بالبحث بأن بياناتهم سيتم التعامل معها بسرية تامة وأنه في حالة نشرها سيتم بذل اقصى الجهود لضمان عدم إمكانية التعرف عليهم من خلالها.
The research team will undertake the followings: • The animal should not be overcome by non-justified burden • Animals should be restrained and transported in a humane manner • Taking care of the animal during the per-operative time with no negligence. • No mutilation of the animal • The animal not euthanatized unless required, with balanced ecosystem • Care for the animal's husbandry. • Care for the infectious, enzootic, epizootic and zoonotic diseases and informing for the notifiable diseases • Disposal of animal's body should be in a proper manner. • Method of killing the used animals.	■	الفريق البحثي يتعهد بمراعاة ما يلي: • عدم تحميل الحيوان بما لا يطيق في كافة الجوانب • الرفق بالحيوان عند التحكم فيه ونقله • الاهتمام بالحيوان قبل وأثناء وبعد إجراء العمليات الجراحية، وعدم إهماله. • عدم التمثيل بالحيوان. • عدم قتل الحيوان إلا لحاجة، ومراعاة التوازن البيئي. • الاهتمام بتربية وتغذية وسياسة الحيوان. • توخي الحرص في التعامل مع الحيوانات بعدم انتشار الأمراض المعدية والمستوطنة والسارية والمتناقلة والإخبار عن الأمراض الخطيرة • التخلص من جثث الحيوانات بالسبل العلمية الصحيحة. • طريقة قتل/ التخلص من الحيوان المستخدم في الدراسة.

Page ٥ of ٦

GNU Free Documentation License

Version 1.2, November 2002

0. Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document *free* in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. Applicability and Definitions

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The “Document”, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as “you”. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A “Modified Version” of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A “Secondary Section” is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document’s overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The “Invariant Sections” are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The “Cover Texts” are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A “Transparent” copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not “Transparent” is called “Opaque”.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The “Title Page” means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, “Title Page” means the text near the most prominent appearance of the work’s title, preceding the beginning of the body of the text.

A section “Entitled XYZ” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “Acknowledgements”, “Dedications”, “Endorsements”, or “History”.) To “Preserve the Title” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. Verbatim Copying

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. Copying in Quantity

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document’s license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. Modifications

you may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties — for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. Combining Documents

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

6. Collections of Documents

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. Aggregation with Independent Works

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. Translation

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, “Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. Termination

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. Future Revisions of this License

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See `http://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License “or any later version” applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © `YEAR YOUR NAME`. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free

> Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled “GNU Free Documentation License”.

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the “with. . . Texts.” line with this:

> with the Invariant Sections being `LIST THEIR TITLES`, with the Front-Cover Texts being `LIST`, and with the Back-Cover Texts being `LIST`.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Bibliography

[1] Christchurch shootings: Social media races to stop attack footage, howpublished = https://www.bbc.co.uk/news/technology-47583393, note = Accessed: Dec 2019.

[2] Hate crime 'police priority' as social media cases soar, howpublished = https://www.bbc.co.uk/news/uk-scotland-glasgow-west-43436900, note = Accessed: May 2018.

[3] Hate speech policy kernel description. https://support.google.com/youtube/answer/2801939. Accessed: June 2019.

[4] Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Accessed: Dec 2017.

[5] Hatespeech. https://www.facebook.com/communitystandards/hate_speech. Accessed: Dec 2018.

[6] Interpreting ROC Curves, Precision-Recall Curves, and AUCs, howpublished = https://www.datascienceblog.net/post/machine-learning/interpreting-roc-curves-auc/, note = Accessed:Nov 2018.

[7] Moderating the internet is hurting workers. How can companies help them?, howpublished = https://edition.cnn.com/2019/02/28/tech/facebook-google-content-moderators/index.html, note = Accessed: July 2019.

[8] Perspective API kernel description. https://www.perspectiveapi.com/.. Accessed: Dec 2017.

[9] Prevent Strategy, howpublished = https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/97976/prevent-strategy-review.pdf, note = Accessed: Aug 2017.

[10] Speech by Commissioner JourovÃ¡ - 10 years of the EU Fundamental Rights Agency: a call to action in defence of fundamental rights, democracy and the rule of law, year=2017.

[11] Swati Agarwal and Ashish Sureka. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442. Springer, 2015.

[12] Swati Agarwal and Ashish Sureka. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv preprint arXiv:1701.04931*, 2017.

[13] Charu C Aggarwal. Similarity and distances. In *Data Mining*, pages 63–91. Springer, 2015.

[14] Kenta Ago, Yoko Uwate, and Yoshifumi Nishio. Influence of local bridge on a complex network of coupled chaotic circuits. In *Proceedings of international Symposium on Nonlinear Theory and its Applications (NOLTAâ14)*, pages 731–734, 2014.

[15] Santa Agreste, Salvatore Catanese, Pasquale De Meo, Emilio Ferrara, and Giacomo Fiumara. Network structure and resilience of mafia syndicates. *Information Sciences*, 351:30–47, 2016.

[16] Sara Ahajjam and Hassan Badir. Identification of influential spreaders in complex networks using hybridrank algorithm. *Scientific reports*, 8(1):1–10, 2018.

[17] Wasim Ahmed. Using twitter as a data source: an overview of social media research tools (2019). https://blogs.lse.ac.uk/impactofsocialsciences/2019/06/18/using-twitter-as-a-data-source-an-overview-of-social-media-research-tools-2 May 2019.

[18] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2016.

[19] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433–443, 2016.

[20] Haider M Al-Khateeb, Gregory Epiphaniou, Zhraa A Alhaboby, James Barnes, and Emma Short. Cyberstalking: Investigating formal intervention and the role of corporate social responsibility. *Telematics and Informatics*, 34(4):339–349, 2017.

[21] Hameed Al-Qaheri and Soumya Banerjee. Measuring homophily in social network: Identification of flow of inspiring influence under new vistas of evolutionary dynamics. In *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Extended Papers from Science and Information Conference*. Citeseer, 2013.

[22] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[23] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378, 2000.

[24] Chris Allen. 10 still a challenge for us all? the runnymede trust, islamophobia and policy. *Religion, Equalities, and Inequalities*, page 113, 2016.

[25] Wafa Alorainy, Pete Burnap, Han Liu, and Matthew Williams. The enemy among us: Detecting hate speech with threats based'othering'language embeddings. *arXiv preprint arXiv:1801.07495*, 2018.

[26] Wafa Alorainy, Pete Burnap, Han Liu, Matthew Williams, and Luca Giommoni. Disrupting networks of hate: characterising hateful networks and removing critical nodes. *Social Network Analysis and Mining*, 12(1):1–22, 2022.

[27] Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L Williams. âthe enemy among usâ detecting cyber hate speech with threats-based othering language embeddings. *ACM Transactions on the Web (TWEB)*, 13(3):1–26, 2019.

[28] Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, 19:100096, 2020.

[29] Dana Alsagheer, Hadi Mansourifar, and Weidong Shi. Counter hate speech in social media: A survey. *arXiv preprint arXiv:2203.03584*, 2022.

[30] Gerhard Andersson, Per Carlbring, and Tomas Furmark. Social anxiety disorder. *The Wiley Blackwell handbook of social anxiety disorder*, page 569, 2014.

[31] Despoina Antonakaki, Paraskevi Fragopoulou, and Sotiris Ioannidis. A survey of twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, 164:114006, 2021.

[32] Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, Maria Qasim, and Imran Ali Khan. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. *PloS one*, 12(2):e0171649, 2017.

[33] Imran Awan. Islamophobia and twitter: A typology of online hate against muslims on social media. *Policy & Internet*, 6(2):133–150, 2014.

[34] Imran Awan and Irene Zempi. âi will blow your face offââvirtual and physical world anti-muslim hate crime. *The British Journal of Criminology*, 57(2):362–380, 2017.

[35] Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, and Adebayo Abayomi-Alli. A probabilistic clustering model for hate speech classification in twitter. *Expert Systems with Applications*, 173:114762, 2021.

[36] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[37] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

[38] Mohit Bansal, Kevin Gimpel, and Karen Livescu. Tailoring continuous word representations for dependency parsing. In *ACL (2)*, pages 809–815, 2014.

[39] Michele Bellingeri, Daniele Bevacqua, Francesco Scotognella, Roberto Alfieri, Quang Nguyen, Daniele Montepietra, and Davide Cassi. Link and node removal in real social networks: A review. *FRONTIERS IN PHYSICS*, 8, 2020.

[40] Michele Bellingeri, Davide Cassi, and Simone Vincenzi. Efficiency of attack strategies on complex model and real-world networks. *Physica A: Statistical Mechanics and its Applications*, 414:174–180, 2014.

[41] Antonio Belmonte, Juan Garrido, Jorge E Jiménez, and Francisco Vázquez. Recomputing causality assignments on lumped process models when adding new simplification assumptions. *Symmetry*, 10(4):102, 2018.

[42] Mario Guajardo-CÃ©spedes Margaret Mitchell Ben Packer, Yoni Halpern. Text embedding models contain bias. here's why that matters, 2018.

[43] Darina Benikova, Michael Wojatzki, and Torsten Zesch. What does this imply? examining the impact of implicitness on the perception of hate speech. In *International Conference of the German Society for Computational Linguistics and Language Technology*, pages 171–179. Springer, 2017.

[44] Kelsey Beninger, Alexandra Fry, Natalie Jago, Hayley Lepps, Laura Nass, and Hannah Silvester. Research using social media; usersâ views. *NatCen Social Research*, pages 1–40, 2014.

[45] Adam Bermingham and Alan F Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836. ACM, 2010.

[46] Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33, 2020.

[47] Irene V Blair, Bernadette Park, and Jonathan Bachelor. Understanding intergroup anxiety: Are some people more anxious than others? *Group processes & intergroup relations*, 6(2):151–169, 2003.

[48] James M Bloodgood, Jeffrey S Hornsby, Matthew Rutherford, and Richard G McFarland. The role of network density and betweenness centrality in diffusing new venture legitimacy: an epidemiological approach. *International Entrepreneurship and Management Journal*, 13(2):525–552, 2017.

[49] Stefano Boccaletti, Ginestra Bianconi, Regino Criado, Charo I Del Genio, Jesús Gómez-Gardenes, Miguel Romance, Irene Sendina-Nadal, Zhen Wang, and Massimiliano Zanin. The structure and dynamics of multilayer networks. *Physics reports*, 544(1):1–122, 2014.

[50] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

[51] Nils Böckler and Thorsten Seeger. Revolution of the dispossessed: School shooters and their devotees on the web. In *School shootings*, pages 309–339. Springer, 2013.

[52] Paolo Boldi, Marco Rosa, and Sebastiano Vigna. Hyperanf: Approximating the neighbourhood function of very large graphs on a budget. In *Proceedings of the 20th international conference on World wide web*, pages 625–634. ACM, 2011.

[53] Paolo Boldi, Marco Rosa, and Sebastiano Vigna. Robustness of social and web graphs to node removal. *Social Network Analysis and Mining*, 3(4):829–842, 2013.

[54] Paolo Boldi and Sebastiano Vigna. Four degrees of separation, really. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1222–1227. IEEE, 2012.

[55] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology*, 2(1):113–120, 1972.

[56] Stephen P Borgatti and Martin G Everett. Network analysis of 2-mode data. *Social networks*, 19(3):243–269, 1997.

[57] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[58] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[59] Annisa Briliani, Budhi Irawan, and Casi Setianingsih. Hate speech detection in indonesian language on instagram comment section using k-nearest neighbor classification method. In *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pages 98–104. IEEE, 2019.

[60] Pete Burnap, Gualtiero Colombo, Rosie Amery, Andrei Hodorog, and Jonathan Scourfield. Multi-class machine classification of suicide-related communication on twitter. *Online social networks and media*, 2:32–44, 2017.

[61] Pete Burnap, Omer F Rana, Nick Avis, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, and Luke Sloan. Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 95:96–108, 2015.

[62] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[63] Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11, 2016.

[64] Pete Burnap, Matthew L Williams, Luke Sloan, Omer Rana, William Housley, Adam Edwards, Vincent Knight, Rob Procter, and Alex Voss. Tweeting the terror: modelling the social media reaction to the woolwich terrorist attack. *Social Network Analysis and Mining*, 4(1):206, 2014.

[65] Peter Burnap and Matthew Leighton Williams. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making. *Internet, Policy Politics, Oxford, UK*, 2014.

[66] Val Burris, Emery Smith, and Ann Strahm. White supremacist networks on the internet. *Sociological focus*, 33(2):215–235, 2000.

[67] Carlos Arcila Calderón, Gonzalo de la Vega, and David Blanco Herrero. Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in spain. *Social Sciences*, 9(11):188, 2020.

[68] Duncan S Callaway, Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468, 2000.

[69] Kathleen M Carley, Ju-Sung Lee, and David Krackhardt. Destabilizing networks. *Connections*, 24(3):79–92, 2002.

[70] Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005.

[71] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[72] Jason Chan, Anindya Ghose, and Robert Seamans. The internet and racial hate crime: Offline spillovers from online access. *MIS Quarterly*, 40(2):381–403, 2016.

[73] Akemi Chatfield and Uuf Brajawidagda. Twitter tsunami early warning network: a social network analysis of twitter information flows. 2012.

[74] Rajdeep Chatterjee, Ankita Datta, and Debarshi Kumar Sanyal. Ensemble learning approach to motor imagery eeg signal classification. In *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*, pages 183–208. Elsevier, 2019.

[75] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Detecting aggressors and bullies on twitter. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 767–768. International World Wide Web Conferences Steering Committee, 2017.

[76] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22. ACM, 2017.

[77] Michael Chau and Jennifer Xu. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, 65(1):57–70, 2007.

[78] Ying Chen. Detecting offensive language in social medias for protection of adolescent online safety. 2011.

[79] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.

[80] Yiping Chen, Gerald Paul, Shlomo Havlin, Fredrik Liljeros, and H Eugene Stanley. Finding a better immunization strategy. *Physical review letters*, 101(5):058701, 2008.

[81] Darko Cherepnalkoski and Igor Mozetič. Retweet networks of the european parliament: evaluation of the community structure. *Applied network science*, 1(1):2, 2016.

[82] Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 2018.

[83] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[84] Reuven Cohen, Keren Erez, Daniel Ben-Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical review letters*, 85(21):4626, 2000.

[85] Reuven Cohen, Keren Erez, Daniel Ben-Avraham, and Shlomo Havlin. Breakdown of the internet under intentional attack. *Physical review letters*, 86(16):3682, 2001.

[86] Samuel K Cohn. Pandemics: waves of disease, waves of hate from the plague of athens to aids. *Historical Research*, 85(230):535–555, 2012.

[87] Andrea Fronzetti Colladon and Peter A Gloor. Measuring the impact of spammers on e-mail and twitter networks. *International Journal of Information Management*, 48:254–262, 2019.

[88] Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications*, 73:291–300, 2016.

[89] Stephen M. Croucher. Integrated threat theory and acceptance of immigrant assimilation: An analysis of muslim immigration in western europe. *Communication Monographs*, 80(1):46–62, 2013.

[90] Paolo Crucitti, Vito Latora, Massimo Marchiori, and Andrea Rapisarda. Efficiency of scale-free networks: error and attack tolerance. *Physica A: Statistical Mechanics and its Applications*, 320:622–642, 2003.

[91] Paolo Crucitti, Vito Latora, Massimo Marchiori, and Andrea Rapisarda. Error and attack tolerance of complex networks. *Physica A: Statistical mechanics and its applications*, 340(1-3):388–394, 2004.

[92] Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 58–65, 2011.

[93] Luigi Curini and Robert Franzese. *The SAGE Handbook of Research Methods in Political Science and International Relations*. Sage, 2020.

[94] Bruno Requião da Cunha and Sebastián Gonçalves. Topology, robustness, and structural controllability of the brazilian federal police criminal intelligence network. *Applied network science*, 3(1):1–20, 2018.

[95] Bruno Requiao da Cunha, Juan Carlos Gonzalez-Avella, and Sebastian Goncalves. Fast fragmentation of networks using module-based attacks. *PloS one*, 10(11), 2015.

[96] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*, 2017.

[97] Milind Dawande, Pinar Keskinocak, Jayashankar M Swaminathan, and Sridhar Tayur. On bipartite and multipartite clique problems. *Journal of Algorithms*, 41(2):388–403, 2001.

[98] Manlio De Domenico, Vincenzo Nicosia, Alexandre Arenas, and Vito Latora. Structural reducibility of multilayer networks. *Nature communications*, 6(1):1–9, 2015.

[99] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.

[100] Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008.

[101] Fabio Del Vigna12, Andrea Cimino23, Felice DellâOrletta, Marinella Petrocchi, and Maurizio Tesconi. Hate me, hate me not: Hate speech detection on facebook. 2017.

[102] Noemi Derzsy. Strategies for combating online hate, 2019.

[103] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18, 2012.

[104] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*, 2011.

[105] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240, 2008.

[106] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, pages 29–30. ACM, 2015.

[107] Margaret E Duffy. Web of hate: A fantasy theme analysis of the rhetorical vision of hate groups online. *Journal of Communication Inquiry*, 27(3):291–312, 2003.

[108] Paul AC Duijn, Victor Kashirin, and Peter MA Sloot. The relative ineffectiveness of criminal network disruption. *Scientific reports*, 4:4238, 2014.

[109] Miles Efron. Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788, 2010.

[110] Mattias Ekman. Anti-immigration and racist discourse in social media. *European Journal of Communication*, 34(6):606–618, 2019.

[111] Louise Ellison and Yaman Akdeniz. Cyber-stalking: the regulation of harassment on the internet. *Criminal Law Review*, 29:29–48, 1998.

[112] Martin Everett and S Borgatti. Extending centrality, pj carrington, j. scott, and s. wasserman, eds, 2005.

[113] Naufal Riza Fatahillah, Pulut Suryati, and Cosmas Haryawan. Implementation of naive bayes classifier algorithm on social media (twitter) to the teaching of indonesian hate speech. In *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, pages 128–131. IEEE, 2017.

[114] Katherine Faust. Comparing social networks: size, density, and local structure. *Metodoloski zvezki*, 3(2):185, 2006.

[115] Leon Festinger, Albert Pepitone, and Theodore M Newcomb. Some consequences of de-individuation in a group. 1963.

[116] Thomas E Ford. Effects of sexist humor on tolerance of sexist events. *Personality and Social Psychology Bulletin*, 26(9):1094–1107, 2000.

[117] Thomas E Ford, Christie F Boxer, Jacob Armstrong, and Jessica R Edel. More than âjust a jokeâ: The prejudice-releasing function of sexist humor. *Personality and Social Psychology Bulletin*, 34(2):159–170, 2008.

[118] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[119] Paula Cristina Teixeira Fortuna. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. 2017.

[120] Abraham H Foxman and Christopher Wolf. *Viral hate: Containing its spread on the Internet*. Macmillan, 2013.

[121] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[122] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[123] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech.* UNESCO Publishing, 2015.

[124] Lazaros K Gallos, Reuven Cohen, Panos Argyrakis, Armin Bunde, and Shlomo Havlin. Stability and topology of scale-free networks under attack and defense strategies. *Physical review letters*, 94(18):188701, 2005.

[125] Björn Gambäck and Utpal Kumar Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, 2017.

[126] Lei Gao, Alexis Kuppersmith, and Ruihong Huang. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. *arXiv preprint arXiv:1710.07394*, 2017.

[127] Phyllis B Gerstenfeld. Hate crime. *The Wiley Handbook of Violence and Aggression*, pages 1–13, 2017.

[128] Phyllis B Gerstenfeld, Diana R Grant, and Chau-Pu Chiang. Hate online: A content analysis of extremist internet sites. *Analyses of social issues and public policy*, 3(1):29–44, 2003.

[129] Sahar Ghannay, Yannick Esteve, and Nathalie Camelin. Word embeddings combination and neural networks for robustness in asr error detection. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 1671–1675. IEEE, 2015.

[130] Manoochehr Ghiassi, James Skinner, and David Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.

[131] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[132] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *2010 Proceedings IEEE Infocom*, pages 1–9. Ieee, 2010.

[133] Jack Glaser, Jay Dixit, and Donald P Green. Studying hate crime with the internet: What makes racists advocate racial violence? *Journal of Social Issues*, 58(1):177–193, 2002.

[134] Benjamin Golub and Matthew O Jackson. Naıve learning in social networks: Convergence, influence, and the wisdom of crowds. *Preprint, available at http://www. stanford. edu/˜ jacksonm/naivelearning. pdf*, 2007.

[135] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[136] Edel Greevy and Alan F Smeaton. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469. ACM, 2004.

[137] Radhouane Guermazi, Mohamed Hammami, and Abdelmajid Ben Hamadou. Using a semi-automatic keyword dictionary for improving violent web site filtering. In *2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, pages 337–344. IEEE, 2007.

[138] Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of all complex networks. *Information processing letters*, 90(5):215–221, 2004.

[139] Ola Hafez. The language and politics of exclusion: Others in discourse: Stephen harold riggins ed., thousand oaks, ca: Sage, 1997. vi+ 294 pp. $25.95(pb.)$,57.95 (hb.), 2000.

[140] Zhang Hailong, Gan Wenyan, and Jiang Bo. Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th web information system and application conference*, pages 262–265. IEEE, 2014.

[141] Steve Halligan, Douglas G Altman, and Susan Mallett. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology*, 25(4):932–939, 2015.

[142] Robert A Hanneman and Mark Riddle. Introduction to social network methods, 2005.

[143] Margaret Harris and Patricia Young. Developing community and social cohesion through grassroots bridge-building: an exploration. *Policy & Politics*, 37(4):517–534, 2009.

[144] James Hawdon, Atte Oksanen, and Pekka Räsänen. Victims of online hate groups. *The causes and consequences of group violence: From bullies to terrorists*, page 165, 2014.

[145] Charlotte Heath-Kelly. Counter-terrorism and the counterfactual: Producing the âradicalisationâdiscourse and the uk prevent strategy. *The British journal of politics and international relations*, 15(3):394–415, 2013.

[146] Itai Himelboim, Kaye D Sweetser, Spencer F Tinkham, Kristen Cameron, Matthew Danelo, and Kate West. Valence-based homophily on twitter: Network analysis of emotions and political talk in the 2012 presidential election. *new media & society*, 18(7):1382–1400, 2016.

[147] Geoffrey E Hinton. Deep belief networks. *Scholarpedia*, 4(5):5947, 2009.

[148] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[149] Bruce Hoffman. *Inside terrorism*. Columbia university press, 2006.

[150] Petter Holme, Beom Jun Kim, Chang No Yoon, and Seung Kee Han. Attack vulnerability of complex networks. *Physical review E*, 65(5):056109, 2002.

[151] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer, 2015.

[152] Mark Hsueh, Kumar Yogeeswaran, and Sanna Malinen. âleave your comment belowâ: Can biased online comments influence our own prejudicial attitudes and behaviors? *Human communication research*, 41(4):557–576, 2015.

[153] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[154] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZC-SRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.

[155] Chung-Yuan Huang, Yu-Hsiang Fu, Yu-Shiuan Tsai, et al. Beyond bond links in complex networks: Local bridges, global bridges and silk links. *Physica A: Statistical Mechanics and its Applications*, 536:121027, 2019.

[156] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International journal of emergency management*, 6(3-4):248–260, 2009.

[157] Fauzia Idrees, Muttukrishnan Rajarajan, Mauro Conti, Thomas M Chen, and Yogachandran Rahulamathavan. Pindroid: A novel android malware detection system using ensemble learning methods. *Computers & Security*, 68:36–46, 2017.

[158] Swami Iyer, Timothy Killingback, Bala Sundaram, and Zhen Wang. Attack robustness and centrality of complex networks. *PloS one*, 8(4), 2013.

[159] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691, 2015.

[160] Ehsan Jahanpour and Xin Chen. Analysis of complex network performance and heuristic node removal strategies. *Communications in Nonlinear Science and Numerical Simulation*, 18(12):3458–3468, 2013.

[161] Almerima Jamakovic and Steve Uhlig. On the relationships between topological measures in real-world networks. *Networks & Heterogeneous Media*, 3(2):345, 2008.

[162] Zubaida Jastania, Rabeeh Ayaz Abbasi, Kawther Saeedi, and Mohammad Ahtisham Aslam. Using social network analysis to understand public discussions: The case study of# saudiwomencandrive on twitter. *INTERNATIONAL JOURNAL*, 11(2):223–231, 2020.

[163] Timothy Jay and Kristin Janschewitz. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288, 2008.

[164] Pablo Jensen, Matteo Morini, Márton Karsai, Tommaso Venturini, Alessandro Vespignani, Mathieu Jacomy, Jean-Philippe Cointet, Pierre Mercklé, and Eric Fleury. Detecting global bridges in networks. *Journal of Complex Networks*, 4(3):319–329, 2016.

[165] Akshita Jha and Radhika Mamidi. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, 2017.

[166] NF Johnson, R Leahy, N Johnson Restrepo, N Velasquez, M Zheng, P Manrique, P Devkota, and Stefan Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261–265, 2019.

[167] Christopher S Josey. Hate speech and identity: An analysis of neo racism and the indexing of identity. *Discourse & Society*, 21(1):27–39, 2010.

[168] Pascal Jürgens, Andreas Jungherr, and Harald Schoen. Small worlds with a difference: New gatekeepers and the filtering of political information on twitter. In *Proceedings of the 3rd international web science conference*, pages 1–5, 2011.

[169] Marcus Kaiser. Mean clustering coefficients: the role of isolated nodes and leafs on clustering measures for small-world networks. *New Journal of Physics*, 10(8):083042, 2008.

[170] Gerald C Kane, Maryam Alavi, Giuseppe Labianca, and Stephen P Borgatti. Whatâs different about social media networks? a framework and research agenda. *MIS quarterly*, 38(1):275–304, 2014.

[171] Andreas M Kaplan and Michael Haenlein. The early bird catches the news: Nine things you should know about micro-blogging. *Business horizons*, 54(2):105–113, 2011.

[172] Mladen Karan and Jan Šnajder. Cross-domain detection of abusive language online. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 132–137, 2018.

[173] Faris Kateb and Jugal Kalita. Classifying short text in social media: Twitter as case study. *International Journal of Computer Applications*, 111(9), 2015.

[174] Irwin Katz. Gordon allport's" the nature of prejudice". *Political Psychology*, pages 125–157, 1991.

[175] Teo Keipi, Matti Näsi, Atte Oksanen, and Pekka Räsänen. *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis, 2016.

[176] Teo Keipi and Atte Oksanen. Self-exploration, anonymity and risks in the online setting: Analysis of narratives by 14–18-year olds. *Journal of Youth Studies*, 17(8):1097–1113, 2014.

[177] Datis Khajeheian. The struggle to regulate social media platforms. *Nordic Journal of Media Management*, 1(2):123–127, 2020.

[178] Kelvin Kiema Kiilu, George Okeyo, Richard Rimiru, and Kennedy Ogada. Using naïve bayes algorithm in detection of hate tweets. *International Journal of Scientific and Research Publications*, 8(3), 2018.

[179] Eunice Kim, Yongjun Sung, and Hamsu Kang. Brand followersâ retweeting behavior on twitter: How brand relationships influence brand electronic word-of-mouth. *Computers in Human Behavior*, 37(37):18–25, 2014.

[180] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[181] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*, 2014.

[182] Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888–893, 2010.

[183] Sebastian Köffer, Dennis M Riehle, Steffen Höhenberger, and Jörg Becker. Discussing the value of automatic hate speech detection in online debates. *Multikonferenz Wirtschaftsinformatik (MKWI 2018): Data Driven X-Turning Data in Value, Leuphana, Germany*, 2018.

[184] Panos Kompatsiaris. Whitewashing the nation: racist jokes and the construction of the african âotherâin greek popular cinema. *Social Identities*, 23(3):360–375, 2017.

[185] Giselinde Kuipers and Barbara Van der Ent. The seriousness of ethnic jokes: Ethnic humor and social change in the netherlands, 1995–2012. *Humor*, 29(4):605–633, 2016.

[186] Jérôme Kunegis. Konect: the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1343–1350. ACM, 2013.

[187] Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *AAAI*, 2013.

[188] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[189] Matthieu Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical computer science*, 407(1-3):458–473, 2008.

[190] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social networks*, 30(1):31–48, 2008.

[191] Matthieu Latapy, Thi Ha Duong Phan, Christophe Crespelle, and Thanh Qui Nguyen. Termination of multipartite graph series arising from complex network modelling. In *International Conference on Combinatorial Optimization and Applications*, pages 1–10. Springer, 2010.

[192] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[193] Elissa Lee and Laura Leets. Persuasive storytelling by hate groups online: Examining its effects on adolescents. *American behavioral scientist*, 45(6):927–957, 2002.

[194] Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou, and Jeffrey Nichols. Who will retweet this? automatically identifying and engaging strangers on

twitter to spread information. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 247–256, 2014.

[195] Laura Leets. Responses to internet hate sites: is speech too free in cyberspace? *Communication Law & Policy*, 6(2):287–317, 2001.

[196] Sune Lehmann and Yong-Yeol Ahn. *Complex spreading phenomena in social systems*. Springer, 2018.

[197] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[198] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[199] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[200] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, 2014.

[201] Ruth Lister. *Poverty and Social Justice: recognition and respect*. Bevan Foundation, 2005.

[202] Han Liu, Pete Burnap, Wafa Alorainy, and Matthew L Williams. A fuzzy approach to text classification with two-stage training for ambiguous instances. *IEEE Transactions on Computational Social Systems*, 6(2):227–240, 2019.

[203] Han Liu, Pete Burnap, Wafa Alorainy, and Matthew L Williams. A fuzzy approach to text classification with two-stage training for ambiguous instances. *IEEE Transactions on Computational Social Systems*, 6(2):227–240, 2019.

[204] Shuhua Liu and Thomas Forss. Combining n-gram based similarity analysis with sentiment analysis in web content classification. In *KDIR*, pages 530–537, 2014.

[205] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

[206] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1–63, 2016.

[207] Pin Luarn and Yu-Ping Chiu. Influence of network density on information diffusion on social network sites: The mediating effects of transmitter activity. *Information Development*, 32(3):389–397, 2016.

[208] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

[209] Clémence Magnien, Matthieu Latapy, and Jean-Loup Guillaume. Impact of random failures and attacks on poisson and power-law random networks. *ACM Computing Surveys (CSUR)*, 43(3):1–31, 2011.

[210] Warih Maharani, Alfian Akbar Gozali, et al. Degree centrality and eigenvector centrality in twitter. In *2014 8th international conference on telecommunication systems services and applications (TSSA)*, pages 1–5. IEEE, 2014.

[211] Arun S Maiya and Tanya Y Berger-Wolf. Benefits of bias: Towards better characterization of network sampling. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 105–113, 2011.

[212] Francesco Martino and Andrea Spoto. Social network analysis: A brief theoretical review and further perspectives in the study of information technology. *PsychNology Journal*, 4(1):53–86, 2006.

[213] Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. Spread of hate speech in online social media. In *Proceedings of the 10th ACM Conference on Web Science*, pages 173–182, 2019.

[214] Tarlach McGonagle et al. The council of europe against online hate speech: Conundrums and challenges. In *Expert paper. Belgrade: Council of Europe Conference of Ministers responsible for Media and Information Society*, 2013.

[215] Lacy G McNamee, Brittany L Peterson, and Jorge Peña. A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs*, 77(2):257–280, 2010.

[216] Priscilla Marie Meddaugh and Jack Kay. Hate speech or 'reasonable racism?' the other in stormfront. *Journal of Mass Media Ethics*, 24(4):251–268, 2009.

[217] Yashar Mehdad and Joel Tetreault. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, 2016.

[218] Andrew Melchionna, Jesus Caloca, Shane Squires, Thomas M Antonsen, Edward Ott, and Michelle Girvan. Impact of imperfect information on network attack. *Physical Review E*, 91(3):032807, 2015.

[219] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[220] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[221] Kristian Miok, Blaž Škrlj, Daniela Zaharie, and Marko Robnik-Šikonja. To ban or not to ban: Bayesian attention networks for reliable hate speech detection. *Cognitive Computation*, pages 1–19, 2021.

[222] Luis Miralles-Pechuán, Dafne Rosso, Fernando Jiménez, and Jose M García. A methodology based on deep learning for advert value calculation in cpm, cpc and cpa networks. *Soft Computing*, 21(3):651–665, 2017.

[223] F Miro-Llinares and JJ Rodriguez-Sala. Cyber hate speech on twitter: Analyzing disruptive events from social media to build a violent communication and hate speech taxonomy. *International Journal of Design & Nature and Ecodynamics*, 11(3):406–415, 2016.

[224] Jozef Miškolci, Lucia Kováčová, and Edita Rigová. Countering hate speech on facebook: The case of the roma minority in slovakia. *Social Science Computer Review*, page 0894439318791786, 2018.

[225] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[226] Rokia Missaoui, Elsa Negre, Dyah Anggraini, and Jean Vaillancourt. Social network restructuring after a node removal. *International journal of Web engineering and technology*, 8(1):4–26, 2013.

[227] Kenny Miyasato. Classification report: Precision, recall, f1-score, accuracy. `https://medium.com/@kennymiyasato/classification-report-precision-recall-f1-score-accuracy-16a245a437a5`, 2016.

[228] Fathali M Moghaddam. The staircase to terrorism: A psychological exploration. *American psychologist*, 60(2):161, 2005.

[229] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94, 2017.

[230] Zewdie Mossie and Jenq-Haur Wang. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087, 2020.

[231] Johann Mourier, Culum Brown, and Serge Planes. Learning and robustness to catch-and-release fishing in a shark social network. *Biology letters*, 13(3):20160824, 2017.

[232] Anna S Mueller and Seth Abrutyn. Suicidal disclosures among friends: using social network data to understand suicide contagion. *Journal of health and social behavior*, 56(1):131–148, 2015.

[233] Kevin Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, 2017.

[234] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.

[235] Ashok Kumar Nanduri, GL Sravanthi, KVKVL Pavan Kumar, Sadhu Ratna Babu, and KVSS Rama Krishna. Modified fuzzy approach to automatic classification of cyber hate speech from the online social networks (osnâs) modified fuzzy approach to automatic classification of cyber hate speech from the online social networks (osnâs).

[236] Matti Johannes Näsi, Pekka Räsänen, Teo Keipi, Atte Oksanen, et al. Trust and victimization: A cross-national comparison of finland, the us, germany and uk. *Research on Finnish society*, 2017.

[237] Hamada A Nayel and HL Shashirekha. Mangalore university inli@ fire2018: Artificial neural network and ensemble based models for inli. In *FIRE (Working Notes)*, pages 110–118, 2018.

[238] Mark Newman. *Networks*. Oxford university press, 2018.

[239] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[240] Mark EJ Newman. The mathematics of networks. *The new palgrave encyclopedia of economics*, 2(2008):1–12, 2008.

[241] Mark EJ Newman, Stephanie Forrest, and Justin Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

[242] Tingyuan Nie, Zheng Guo, Kun Zhao, and Zhe-Ming Lu. New attack strategies for complex networks. *Physica A: Statistical Mechanics and its Applications*, 424:248–253, 2015.

[243] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[244] John T Nockleby. Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279, 2000.

[245] Atte Oksanen, James Hawdon, Emma Holkeri, Matti Näsi, and Pekka Räsänen. Exposure to online hate among young social media users. In *Soul of society: a focus on the lives of children & youth*. Emerald Group Publishing Limited, 2014.

[246] Dion RJ O'Neale et al. Degree distributions of bipartite networks and their projections. *arXiv preprint arXiv:1802.04953*, 2018.

[247] Oluwafemi Oriola and Eduan Kotzé. Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. *IEEE Access*, 8:21496–21509, 2020.

[248] Momoko Otsuka and Sho Tsugawa. Robustness of network attack strategies against node sampling and link errors. *PloS one*, 14(9), 2019.

[249] Sefa Ozalp, Matthew L Williams, Pete Burnap, Han Liu, and Mohamed Mostafa. Antisemitism on twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media+ Society*, 6(2):2056305120916850, 2020.

[250] Anil Özberk and İlyas Çiçekli. Offensive language detection in turkish tweets with bert models. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 517–521. IEEE, 2021.

[251] Alessio Pagani, Guillem Mosquera, Aseel Alturki, Samuel Johnson, Stephen Jarvis, Alan Wilson, Weisi Guo, and Liz Varga. Resilience or robustness: identifying topological vulnerabilities in rail networks. *Royal Society open science*, 6(2):181301, 2019.

[252] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[253] Christopher R Palmer, Georgos Siganos, Michalis Faloutsos, Christos Faloutsos, and Phillip B Gibbons. The connectivity and fault-tolerance of the internet topology. 2001.

[254] Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, pages 363–370, 2019.

[255] Raj Kumar Pan and Jari Saramäki. The strength of strong ties in scientific collaboration networks. *EPL (Europhysics Letters)*, 97(1):18007, 2012.

[256] Aasish Pappu and Amanda Stent. Location-based recommendations using nearest neighbors in a locality sensitive hashing (lsh) index.(nov. 20 2015). *US Patent App*, 14(948,213), 2015.

[257] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*, 2017.

[258] Juan Manuel Pastor, Silvia Santamaría, Marcos Méndez, and Javier Galeano. Effects of topology on robustness in ecological bipartite networks. *Networks & Heterogeneous Media*, 7(3):429, 2012.

[259] Romualdo Pastor-Satorras and Alessandro Vespignani. Immunization of complex networks. *Physical review E*, 65(3):036104, 2002.

[260] Indraneil Paul, Abhinav Khattar, Ponnurangam Kumaraguru, Manish Gupta, and Shaan Chopra. Elites tweet? characterizing the twitter verified user network. *arXiv preprint arXiv:1812.09710*, 2018.

[261] Etienne Pelaprat and Barry Brown. Reciprocity: Understanding online social relations. *First Monday*, 17(10), 2012.

[262] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[263] Barbara Perry and Patrik Olsson. Cyberhate: the globalization of hate. *Information & Communications Technology Law*, 18(2):185–199, 2009.

[264] Rasmus Rosenqvist Petersen, Christopher J Rhodes, and Uffe Kock Wiil. Node removal in criminal networks. In *2011 European Intelligence and Security Informatics Conference*, pages 360–365. IEEE, 2011.

[265] K Pfeffers, Tuure Tuunanen, Charles E Gengler, Matti Rossi, Wendy Hui, Ville Virtanen, and Johanna Bragge. The design science research process: A model for producing and presenting information systems research. In *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006), Claremont, CA, USA*, pages 83–106, 2006.

[266] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48(12):4730–4742, 2018.

[267] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[268] Robert D Putnam et al. *Bowling alone: The collapse and revival of American community*. Simon and schuster, 2000.

[269] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[270] A Ramachandra Rao and Suraj Bandyopadhyay. Measures of reciprocity in a social network. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 141–188, 1987.

[271] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015.

[272] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. " like sheep among wolves": Characterizing hateful users on twitter. *arXiv preprint arXiv:1801.00317*, 2017.

[273] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. *arXiv preprint arXiv:1803.08977*, 2018.

[274] Axel Rodriguez, Carlos Argueta, and Yi-Ling Chen. Automatic detection of hate speech on facebook using sentiment and emotion analysis. In *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 169–174. IEEE, 2019.

[275] Damian Roland, Jesse Spurr, and Daniel Cabrera. Preliminary evidence for the emergence of a health care online community of practice: using a netnographic framework for twitter hashtag analytics. *Journal of medical Internet research*, 19(7):e252, 2017.

[276] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

[277] Eldar Sadikov and Maria Montserrat Medina Martinez. Information propagation on twitter. *CS322 project report*, 2009.

[278] Raazesh Sainudiin, Kumar Yogeeswaran, Kyle Nash, and Rania Sahioun. Characterizing the twitter network of prominent politicians and splc-defined hate groups in the 2016 us presidential election. *Social Network Analysis and Mining*, 9(1):34, 2019.

[279] Joni Salminen, Maximilian Hopf, Shammur A Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J Jansen. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1–34, 2020.

[280] Kl Saravanan and S Sasithra. Review on classification based on artificial neural networks. *International Journal of Ambient Systems and Applications (IJASA)*, 2(4):11–18, 2014.

[281] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain*, pages 1–10, 2017.

[282] Christian M Schneider, Tamara Mihaljev, and Hans J Herrmann. Inverse targetingâan effective immunization strategy. *EPL (Europhysics Letters)*, 98(4):46002, 2012.

[283] John Scott. Social network analysis. *Sociology*, 22(1):109–127, 1988.

[284] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[285] Jacob S Siegel. *Demographic and socioeconomic basis of ethnolinguistics*. Springer, 2018.

[286] Scharolta Katharina Sienčnik. Adapting word2vec to named entity recognition. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, number 109, pages 239–243. Linköping University Electronic Press, 2015.

[287] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. In *ICWSM*, pages 687–690, 2016.

[288] Ashley G Smart, Luis AN Amaral, and Julio M Ottino. Cascading failure and robustness in metabolic networks. *Proceedings of the National Academy of Sciences*, 105(36):13223–13228, 2008.

[289] Avocado social. The latest uk social media statistics for 2018. `https://avocadosocial.com/the-latest-uk-social-media-statistics-for-2018/`, May 2018.

[290] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, 44(2):136–146, 2018.

[291] Raymond T Sparrowe, Robert C Liden, Sandy J Wayne, and Maria L Kraimer. Social networks and the performance of individuals and groups. *Academy of management journal*, 44(2):316–325, 2001.

[292] Amedapu Srinivas and R Leela Velusamy. Identification of influential nodes from social networks based on enhanced degree centrality measure. In *2015 IEEE International Advance Computing Conference (IACC)*, pages 1179–1184. IEEE, 2015.

[293] Ervin Staub. The origins and evolution of hate, with notes on prevention. 2005.

[294] Karen Stepanyan, Kerstin Borau, and Carsten Ullrich. A social network analysis perspective on student interaction within the twitter microblogging environment. In *2010 10th IEEE International Conference on Advanced Learning Technologies*, pages 70–72. IEEE, 2010.

[295] Walter G Stephan and Cookie White Stephan. Intergroup threat theory. *The International Encyclopedia of Intercultural Communication*, 2009.

[296] Walter G Stephan and Cookie White Stephan. Intergroup threat theory. *The International Encyclopedia of Intercultural Communication*, pages 1–12, 2017.

[297] Walter G Stephan, Cookie White Stephan, and William B Gudykunst. Anxiety in intergroup relations: A comparison of anxiety/uncertainty management theory and integrated threat theory. *International Journal of Intercultural Relations*, 23(4):613–628, 1999.

[298] Walter S Stephan and Cookie White Stephan. An integrated threat theory of prejudice. In *Reducing prejudice and discrimination*, pages 33–56. Psychology Press, 2013.

[299] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social mediaâsentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013.

[300] Rebecca Stotzer. Comparison of hate crime rates across protected and unprotected groups. *The Williams Institute*, 2007.

[301] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184. IEEE, 2010.

[302] Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, 2019.

[303] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[304] Luke Kien-Weng Tan, Jin-Cheon Na, Yin-Leng Theng, and Kuiyu Chang. Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration. *Journal of Computer Science and Technology*, 27(3):650–666, 2012.

[305] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Education India, 2016.

[306] Natalya Tarasova. Classification of hate tweets and their reasons using svm, 2016.

[307] Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160, 1972.

[308] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[309] Neil Thompson. *Anti-discriminatory practice: Equality, diversity and social justice*. Macmillan International Higher Education, 2016.

[310] I-Hsien Ting, Shyue-Liang Wang, Hsing-Miao Chi, and Jyun-Sing Wu. Content matters: A study of hate groups detection based on social networks analysis and web mining. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 1196–1201. IEEE, 2013.

[311] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*. Citeseer, 2010.

[312] Brendesha M Tynes, Juan Del Toro, and Fantasy T Lozada. An unwelcomed digital visitor in the classroom: The longitudinal impact of online racial discrimination on academic motivation. *School psychology review*, 44(4):407–424, 2015.

[313] Teun A Van Dijk. *Elite discourse and racism*, volume 6. Sage, 1993.

[314] Tin Van Huynh, Vu Duc Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. Hate speech detection on vietnamese social media text using the bi-gru-lstm-cnn model. *arXiv preprint arXiv:1911.03644*, 2019.

[315] Demival Vasques Filho and Dion RJ O'Neale. Degree distributions of bipartite networks and their projections. *Physical Review E*, 98(2):022307, 2018.

[316] Svitlana Volkova and Eric Bell. Identifying effective signals to predict deleted and suspended accounts on twitter across languages. In *ICWSM*, pages 290–298, 2017.

[317] Pooja Wadhwa and MPS Bhatia. Discovering hidden networks in on-line social networks. *International Journal of Intelligent Systems and Applications*, 6(5):44–54, 2014.

[318] JW Wang, LL Rong, and TZ Guo. A new measure method of network node importance based on local characteristics. *Journal of Dalian University of Technology*, 50(5):822–826, 2010.

[319] Xiao Fan Wang and Guanrong Chen. Complex networks: small-world, scale-free and beyond. *IEEE circuits and systems magazine*, 3(1):6–20, 2003.

[320] DOUGLAS WARFIELD. Is/it research: A research methodologies review. *Journal of theoretical & applied information technology*, 13, 2010.

[321] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.

[322] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: a typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.

[323] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of NAACL-HLT*, pages 88–93, 2016.

[324] Stanley Wasserman, Katherine Faust, et al. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[325] Duncan J Watts and Steven H Strogatz. Collective dynamics of âsmall-worldânetworks. *nature*, 393(6684):440, 1998.

[326] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631. ACM, 2005.

[327] Uffe Kock Wiil, Jolanta Gniadek, and Nasrullah Memon. Measuring link importance in terrorist networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 225–232. IEEE, 2010.

[328] Quintan Wiktorowicz and Shahed Amanullah. How tech can fight extremism. *CNN. http://edition. cnn. com/2015/02/16/opinion/wiktorowicz-tech-fighting-extremism/index. html*, 2015.

[329] Matthew Williams. The connection between online hate speech and real-world hate crime. *Retrieved February*, 11:2020, 2019.

[330] Matthew Williams. The science of hate: How prejudice becomes hate and what we can do to stop it, 2021.

[331] Matthew L Williams and Pete Burnap. Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2):211–238, 2016.

[332] Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 2019.

[333] Matthew L Williams, Pete Burnap, and Luke Sloan. Towards an ethical framework for publishing twitter data in social research: Taking into account usersâ views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168, 2017.

[334] Matthew L Williams and Jasmin Tregidga. Hate crime victimization in wales: Psychological and physical impacts across seven hate crime victim types. *British Journal of Criminology*, 54(5):946–967, 2014.

[335] Rob Williams and David Lusseau. A killer whale social network is vulnerable to targeted removals. *Biology letters*, 2(4):497–500, 2006.

[336] Ruth Wodak. *Discursive construction of national identity*. Edinburgh University Press, 2009.

[337] Ruth Wodak and Norman Fairclough. *Critical discourse analysis*. Sage, London, 1997.

[338] Ruth Wodak and Martin Reisigl. Discourse and racism: European perspectives. *Annual Review of Anthropology*, 28(1):175–199, 1999.

[339] Jeremy Wright and Sajid Javid. Online harms white paper. april 2019. 2019.

[340] Ying Wu and Zhiguang Duan. Social network analysis of international scientific collaboration on psychiatry research. *International journal of mental health systems*, 9(1):1–10, 2015.

[341] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee, 2017.

[342] Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM, 2012.

[343] Jennifer Xu and Hsinchun Chen. The topology of dark networks. *Communications of the ACM*, 51(10):58–65, 2008.

[344] Qi Xuan, Yewei Yu, Jinyin Chen, Zhongyuan Ruan, Zhong Fu, and Zicong Lv. Robustness analysis of bipartite task assignment networks: A case study in hospital logistics system. *IEEE Access*, 7:58484–58494, 2019.

[345] Yang Yang, Takashi Nishikawa, and Adilson E Motter. Small vulnerable sets determine large network cascades in power grids. *Science*, 358(6365), 2017.

[346] Mihalis Yannakakis. Node-deletion problems on bipartite graphs. *SIAM Journal on Computing*, 10(2):310–327, 1981.

[347] Michele L Ybarra, Marie Diener-West, Dana Markow, Philip J Leaf, Merle Hamburger, and Paul Boxer. Linkages between internet and other media violence with seriously violent behavior by youth. *Pediatrics*, 122(5):929–937, 2008.

[348] Michele L Ybarra, Kimberly J Mitchell, and Josephine D Korchmaros. National trends in exposure to and experiences of violence on the internet among children. *Pediatrics*, 128(6):e1376–e1386, 2011.

[349] Ruan Yi-Run, Lao Song-Yang, Wang Jun-De, Bai Liang, and Chen Li-Dong. Node importance measurement based on neighborhood similarity in complex network. *Acta Physica Sinica*, 66(3), 2017.

[350] Wenjie Yin and Arkaitz Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 2021.

[351] Michael Yip, Nigel Shadbolt, and Craig Webber. Structural analysis of online criminal social networks. In *2012 IEEE International Conference on Intelligence and Security Informatics*, pages 60–65. IEEE, 2012.

[352] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89, 2011.

[353] Peng Zhang, Jinliang Wang, Xiaojia Li, Menghui Li, Zengru Di, and Ying Fan. Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 387(27):6869–6875, 2008.

[354] Ziqi Zhang and Lei Luo. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *arXiv preprint arXiv:1803.03662*, 2018.

[355] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer, 2018.

[356] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, Heraklion, Crete, 3-7 June 2018.

[357] Xiaohui Zhao, Fangâai Liu, Jinlong Wang, Tianlai Li, et al. Evaluating influential nodes in social networks by local centrality with a coefficient. *ISPRS International Journal of Geo-Information*, 6(2):35, 2017.

[358] Dengyong Zhou, Bernhard Schölkopf, and Thomas Hofmann. Semi-supervised learning on directed graphs. In *NIPS*, volume 5, pages 1633–1640, 2004.

[359] Yilu Zhou, Edna Reid, Jialun Qin, Hsinchun Chen, and Guanpi Lai. Us domestic extremist groups on the web: link and content analysis. *IEEE intelligent systems*, 20(5):44–51, 2005.

[360] Aleksandr Aleksandrovich Zykov. *Fundamentals of graph theory*. BCS Associates Moscow, 1990.