4-27-2023

A Graph-Based Approach to Studying the Spread of Radical Online Sentiment

Le Nguyen
ln8378@rit.edu

A Graph-Based Approach to Studying the Spread of Radical Online Sentiment

by

Le Nguyen

A Thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in Data Science
Department of Software Engineering
B. Thomas Golisano College of Computing and Information Sciences

Rochester Institute of Technology
Rochester, New York
April 27th, 2023

Graph-Based Approach to Studying the Spread of Radical Online Sentiment

Le Nguyen

Committee Approval:

We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Master's in Data Science.

Dr. Nidhi Rastogi Dissertation Advisor	Date
Dr. Ashique KhudaBukhsh Dissertation Committee Member	Date
Dr. Andy Meneely Dissertation Committee Member	Date
Dr. Mohamed Wiem Mkaouer Dissertation Committee Member	Date
Dr. Travis Desell Graduate Program Director, Data Science	Date

Abstract

The spread of radicalization and extremism through the Internet is a growing problem. We are witnessing a rise in online hate groups, inspiring the impressionable and vulnerable population towards extreme actions in the real world. Though the body of research to address these issues is growing in kind, they lack a key understanding of the structure and behavior of online extremist communities. In this thesis, we study the structure and behavior of extremist online communities and the spread of hateful sentiments through them to address this gap in the research. We propose a novel Graph-Based Approach to Studying the Spread of Radical Online Sentiment for studying the dynamics of online comment threads by representing them as graphs. Our Graph-Based Approach to Studying the Spread of Radical Online Sentiment allows us to leverage network analysis tools to reveal the most influential members in a social network and investigate sentiment propagation through the network as well. By combining sentiment analysis, social network analysis, and graph theory, we aim to shed light on the propagation of hate speech in online forums and the extent to which such speech can influence individuals.

In this thesis, we pose four main research questions; firstly, *to what extent do connected members in an online comment thread and connected threads themselves share sentiment? Further, what is the impact of the frequency of interaction, measured by the degree of connection, on the sharing of sentiment?*. Secondly, *who are the most influential members in a comment thread, and how do they shape the sentiment in that thread?* Thirdly, *what does the sentiment of the thread look like over time as more members join threads and more comments are made?* Finally, *can the behavior of online sentiment spread be generalized? Can we develop a model for it?* To answer these questions, we apply our Graph-Based Approach to Studying the Spread of Radical Online Sentiment to 1,973 long comment threads (30+ comments), totaling to 137k comments posted on dark-web forums. These threads contain a combination of benign posts and extremist comments on the Islamic religion from an unmoderated source. To

answer our first research question, we constructed intra- and inter-thread graphs where we could analyze weighted and unweighted connections between threads and members within threads. Our results show that 73% of connected members within a comment thread shares a similar sentiment, and 64% of connected comment threads share a similar sentiment on the inter-thread level when weighted by the degree of connection. Additionally, we found the most influential members of our graphs using information centrality. We found that the original poster was the most influential member in our comment threads 57% of the time, with the mean sentiment of the thread matching the sentiment of the original poster. For our third research question, we performed a temporal analysis of our threads. This analysis further supported our findings in our second research question. Over time, the majority of our threads had their overall sentiment regress to the sentiment of the original poster, with the original poster being the member with the highest influence for 40% of the time steps.

For our fourth and final research question, we used our understanding of our comment threads to create a model that can classify the sentiment of a thread member based on the members they are connected with. We achieved 87% accuracy with our classification model and further used it as a sentiment contagion model, which predicted the sentiment of a new member to a thread based on existing members with 72% accuracy. We plan to expand our study and further the robustness of our models on larger data sets and incorporate stance detection tools.

The complete code is available at https://github.com/nguye639/NguyenThesis

Research Products

- Le Nguyen and Nidhi Rastogi, Graph-based Approach for Studying Spread of Radical Online Sentiment, Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA, https://dl.acm.org/doi/10.1145/3543873.3587634.

- Le Nguyen, Graph-based Approach for Studying Spread of Radical Online Sentiment, Presentation at UpStat 2023, April 22, 2023, Rochester, New York, https://community.amstat.org/rochester/events/upstats-2023. First Place Winner of Applied Statistics Category.

Acknowledgements

This work was done under the mentoring of Dr. Nidhi Rastogi, Principal Investigator, and committee members Dr. Andy Meneenly, Dr.Ashique KhudaBukhsh, and Dr. Mohammed Wiem Mkaouer. I would like to thank all of them for their time and guidance through the thesis process. Though our time together was short, we made the most of it. This work would have never gone as far as it did without them.

This material is based upon work supported by the Graduate Fellowship for STEM Diversity funding (GFSD). I also need to give thanks to Dr. Derek Armstrong and Dr. Eric Nelson for nominating me for the GFSD through Los Alamos National Laboratory. Without the financial support of this fellowship, this work would have never happened.

To my dearest brother Pham and his wife Kayla

Congratulations on their newborn son, my nephew, Anh Porter Nguyen

Contents

1 Introduction

Through the Covid-19 pandemic, there has been an increase in online activity, including the spread of hateful and violent ideologies. Though the presence of online hate is not a new phenomenon, with more of life shifted to the online world, the Internet is now a significant vector for the spread of hateful rhetoric [1,2]. The anonymity and ease of access to the Internet has allowed the spread of this harmful rhetoric with minimal accountability to those responsible but not to the world at large. This has led to increased online hate towards minorities and incidents of terror attacks targeting marginalized communities. These incidents can have devastating consequences and must be addressed to stem severe real-world consequences [1, 3]. Online Hate Research (OHR) has grown in response to these trends. A literature survey by Waqas et al. [4] found that the number of publications in this field has rapidly increased, with a 1000% increase between 2005 and 2018. This work is heavily focused on using natural language processing (NLP) to develop models for the classification of extremist speech and the prediction of what speech will become extreme with the intent to stem its effects. Models used in such research focus on language characteristics such as sentiment and semantics to determine what speech is extreme or what speech will become extreme [5]. As extensive as this work is, it is limited to looking at the extremist language used online and not the dynamics of online extremist communities themselves [6].

The lack of interest in community dynamics when studying online extremist language is a gap in the research. We should not only be looking into how to classify and predict extremist language but also look into how it spreads and how it is adopted in a community [7,8]. Understanding the community dynamics and social contagion of extremism will not only help stem its effects, but it will also contribute to the main body of research by adding an extra feature space to feed into classification and prediction models.

In this work, we expand our understanding of online extremist communities by studying their communication dynamics and how extremist language spreads.

We use natural language processing and social network analysis to analyze the propagation and contagion of sentiment, the most significant feature in extremist language classifiers, to do this [5]. This analysis is done within comment threads as well as between comment threads so we can understand the micro and macro behaviors of extremist communities as well as the interplay between both levels. Our work aims to provide valuable insights into the dynamics of the spread of extremism online in hopes of informing the development of strategies to mitigate its spread.

1.1 Research Questions

RQ1: To what extent do connected members in an online comment thread and connected threads themselves share sentiment? Further, what is the impact of the frequency of interaction, measured by the degree of connection, on the sharing of sentiment?

RQ2: Who are the most influential members in a comment thread, and how do they shape the sentiment in that thread?

RQ3: What does the sentiment of the thread look like over time as more members join threads and more comments are made?

RQ4: Can the behavior of online sentiment spread be generalized? Can we develop a model for it?

1.2 Scope

Every component of this work is an active area of research in itself; from sentiment analysis to NLP on radical speech and even graph theory. We will describe all of the disciplines we will be using and state whether or not this work will expand on them.

1. Graph/Network Application: Graph theory provides a powerful tool for modeling and analyzing complex systems of connections, such as social networks or online communities. By representing individuals or groups

as nodes in a graph, and their relationships (or links) as edges, we can gain insights into the structure of the network, identify key influencers or connectors, and study the flow of information and influence through the network. This work will focus on applying graph theory to the study of online communities to understand how sentiments propagate and how radicalization occurs within online communities.

2. Sentiment Analysis: Sentiment analysis is a significant field in NLP. It allows us to computationally determine the connotation of text and speech for further analysis. The advent of large language models has given us the ability to determine the emotion of text within the context with extraordinary accuracy. The work in this thesis heavily relies on sentiment analysis as we are concerned with how the sentiment propagates through online comment threads. We will be using a state-of-the-art large language model for our sentiment analysis and trust that it is giving us accurate results. This work is not interested in expanding or further refining the field of sentiment analysis, simply using the best sentiment analysis tools available.

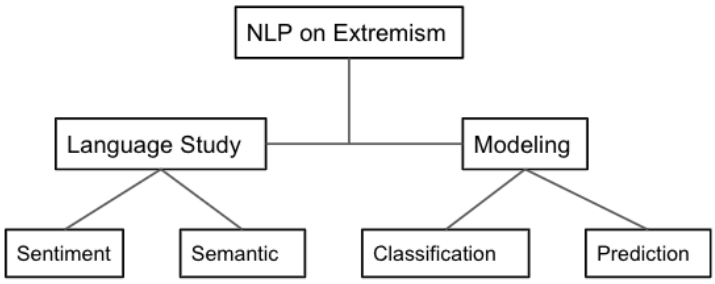


Figure 1: Landscape of NLP on Extremism

3. Study of Extremist Language: The study of extremist language extends to many domains, such as linguistics, political science, and sociology. These domains are interested in understanding what is extremist language, what makes language extreme, and its properties and effects. In the field of On-

line Hate Research, they use the findings from those domains and NLP to make models that can classify extremist speech and predict what speech will become extreme [6]. In our work, we will use and accept the findings from prior research on extremist language, but we will not pursue any language study of extremist speech ourselves. For instance, we will be accepting a definition of what extremist speech is from literature: *extremism* refers to an anti-democratic movement and stands against "all those who do not embrace its dogmatic recipe for a transformation of society." [6]. We are interested in the spread of extremist speech but not the linguistic properties of the speech itself.

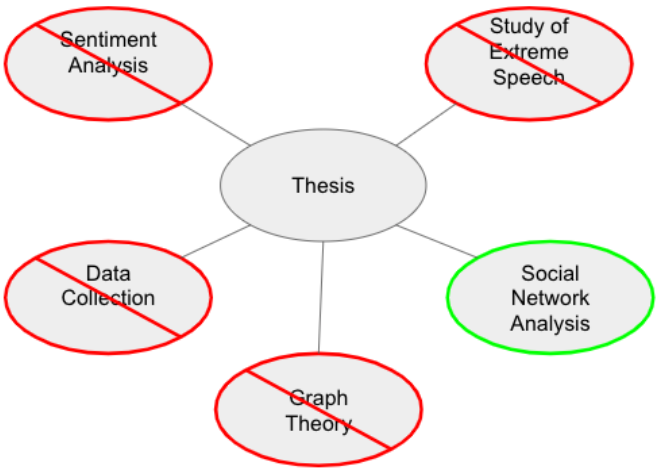


Figure 2: Scope of Thesis

1.3 Scientific Merit

The research done in this thesis has several scientific merits. Firstly, it addresses the urgent and pressing issue of online radicalization and extremism. Online radicalization and extremism are a growing concern in the Covid-19 era and onward, as we have seen an uptick in terrorist events caused by such online activity [9]. By examining how hateful and extremist sentiment spreads through

online forums, this thesis aims to shed light on the mechanisms of radicalization. This research can then be used to identify interventions to prevent radicalization and the spread of extremist ideology.

Secondly, the study employs a novel approach that utilizes graph-based methods to investigate the spread of extremist sentiments in online communities. It expands the application of social network analysis and social contagion modeling. These novel applications further the utility of such methods and strengthen their validity.

Thirdly, this study has broader implications for NLP and social network analysis. The proposed research aims to develop new methods for analyzing sentiment spread in online forums, which may have applications in various fields, from market research to political analysis. This work can later be adapted to study other classes of sentiment beyond hateful and extreme, which will allow researchers to study other social phenomena.

1.4 Broader Impact

This work will make a significant impact on the study of online extremism and hopes to stem its future impact. The research sheds light on the mechanisms of radicalization and can be used to identify potential interventions to prevent it and inform policy decisions in countering the spread of extremist ideologies online. In addition, the models developed in this research that combine graph and sentiment analysis can be used in contexts beyond studying the spread of radicalization. For instance, the graph-based SNA approach can later be repurposed to study the spread of misinformation, propaganda, or the propagation of anything else through the internet. Several other fields, such as political science and journalism, where understanding the spread of information is critical, can benefit from this research.

1.5 Thesis Overview

The rest of this thesis will go about answering research questions 1-4 in the following structure:

- We will first give a thorough background on the topic of this thesis to contextualize the work within the research domain as well as give definitions for all of the terminology we are using.
- We lay out the related work in our domain with an in-depth literature review and analysis of existing solutions to our problem.
- We then give our proposed approach and describe our Graph-Based Approach to Studying the Spread of Radical Online Sentiment.
- We state the results of our experiments and give our evaluation. First, describing the data set we experimented on, then broke down our results, and finally explored work that went beyond our research questions.
- Finally, we analyze our work, discussing the work's limitations and what future work we wish to perform. The thesis ends with a conclusion summarizing the work in its entirety.

2 Background

In this section, we will be laying out all of the background knowledge and concepts needed to understand the work done in this thesis.

2.1 Contextualization of Work Within the Research Domain

The domain of using NLP to study online hate is very expansive. The number of contributing publications has been exponentially increasing, generating many literature reviews [4,6,10]. The work in this field can generally be split into tool and technique papers which develop the technical NLP side, and then language

and topic study papers that apply the NLP tools to analyze radical speech [4,6]. Our work is solidly in the former category but will have to take guidance from the latter to make an effective tool.

The language papers mainly focus on Jihadist terrorism, but a growing number of them are starting to analyze far-right extremism [6]. Within our domain, the data set we are using (AZSecure Dark Web Forums [11]) is appropriate as it pertains to Islamic extremism and has been used in other relevant work [12,13].

In terms of tools and techniques, there is a significant focus on the classification of radical speech, which includes developing natural language feature extraction techniques [4, 6, 10] and a minor focus on predicting if/what speech will become radical which still uses classification and language features [6,13–15]. Our work falls more into the prediction category but takes a different approach than the current body of work. At this point, we are not looking at what language markers can be used to predict radicalization; we are looking at how radicalization can spread using sentiment as an indicator.

After looking at several recent literature surveys [4, 6, 10], our work is novel and relevant to the current interests of the field. We can develop a tool to study how radicalization spreads which may give insight into how to classify radical speech (looking at how a given agent is connected to others can be a feature). Furthermore, we can predict how said radical speech will spread after generalizing the behavior, giving insight into what/how speech becomes radical.

2.2 Definitions

This section is provided for readers to understand the specific vernacular of this thesis as well as expose them to concepts before they are discussed.

- Information Diffusion: This refers to spreading information through a population. Information spreads through a population by members interacting in various ways, whether in person, online, in writing, or in speech. Information diffusion can describe the spread of rumors, the effectiveness of an advertising campaign, and any other information that can spread

through a population.

- Social Contagion: This refers to how information spreads through a population in the diffusion process and the various factors that influence it, such as social connections, media coverage, and cultural norms. Social contagion considers how fast information spreads (how contagious it is), how much of a population the information spreads to, and the resistance members of a population can give to new information. Examples include the rapid spread of a viral video, the propagation of conspiracy theories through online communities, or the widespread adoption of a new trend.

- Natural Language Processing (NLP): This refers to the field of study and set of tools that allow computers to understand written and spoken language the same way humans can. Examples of natural language processing include sentiment analysis, classification of text, predictive text generation, and part of speech tagging.

- Sentiment Analysis: This refers to computationally determining the emotion conveyed in language. Specifically, computational tools can be used to determine the emotion in written or spoken language. Historically this has been done lexically with dictionaries full of words and their given sentiment, but with the rise of large language models, they are the current go-to method. Sentiment analysis can be single-class and give how positive or negative a text is, or multi-class, giving the set of emotions and the degree they are conveyed. Uses of sentiment analysis include automated analysis of customer feedback and advertising impact.

- Graphs, Nodes, & Edges: A graph is a mathematical representation of information that shows the value of data points and their relationship. A node represents every data point, and they are connected by edges that show the relationship between points. Graphs can represent the relationships between people, the roads connecting two buildings, and all the possible moves on a chess board.

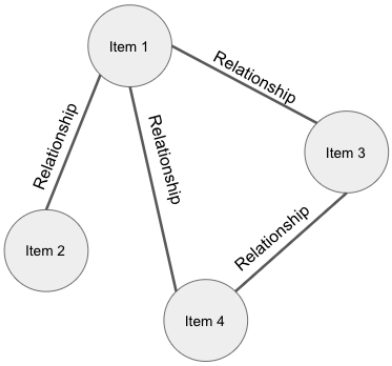


Figure 3: Example Graph

- Centrality: A measurement of the most central member in a graph. Many centrality metrics exist, but a commonly used one is degree centrality which determines the most central node to be the one with the most edges to other nodes. Centrality can be used to find the most influential person in a social network, which roads will have the most traffic, and which web pages are most relevant to a search.

- Degree of Connection: A measurement of how connected two nodes are in a graph. Usually, this is done with weighted edges where the weight of the edge represents the degree of connection.

- Social Network Analysis (SNA): This refers to the methods used for studying the relationships between members of a community or network, using graph theory to model the social structure. In SNA, nodes represent individuals or groups, and edges represent the connections between them (such as friendships, professional relationships, or shared interests). Analyzing patterns of connections in a social network can help gain insights into the influence of information and flow through the network.

- Inter-thread: This refers to studying the relationships between comment threads. In inter-thread analysis, comment threads are represented by nodes connected by edges showing they share commenters.

- Intra-thread: This refers to studying the relationships inside of comment threads. In intra-thread analysis, commenters inside of threads are represented by nodes connected by edges to the other commenters they replied to.

- Extremism: This refers to anti-democratic movements that stand against any that do not embrace their dogmatic changes to society. Examples of extremism include Jihadism, terrorism, and white supremacy.

- Radicalization: This refers to the process of causing someone to adopt radical or contrarian ideas. Radicalization can be done through political messaging, disinformation campaigns, and acts of terrorism.

- Dark Web: This refers to the portion of the internet not indexed by standard search engines and is often used for illegal or illicit activities such as drug trafficking, weapons sales, or child pornography. The Dark Web is accessed through specialized browsers that allow users to browse anonymously and is often associated with underground marketplaces and online communities where criminal activity is facilitated [16].

3 Literature Review and Related Work

In this section, we go over the relevant literature in our domain as well as work related to the work done in this thesis.

3.1 How Context Impacts Sentiment Analysis

The impact of context on sentiment has been thoroughly discussed in NLP work [5,6,10,17,18]. When a word is being used, who is using a word, and what is being referred to are all things to consider when looking at the sentiment of a word [17]. For instance, the word *soft* can have different connotations depending on the context it is used in. Calling a person soft is an insult, but referring to an animal or a tone as *soft* is a compliment [17].

When looking at the language used in our data set, we need to keep this in mind since we are looking at the language of a specific community. Extremist Jihadist language has its vocabulary and connotes differently than regular speech. One of the best examples is the phrase *allahu akbar* (God is great), used in praise and prayer. In general, praising God would have a positive connotation, but Jihadist language is used to praise horrific acts [19].

We must also consider that many posters in our data set are non-native English speakers and will bring in language traits from their native tongue, primarily Arabic. This means we must deal with atypical phrasing and vocabulary usage, which may break standard NLP tools built on English data sets [17].

3.2 Connection Between Radicalization and Sentiment

A link between sentiment and radicalization has been established in literature [5, 6, 10]. Both single-class and multi-class sentiment scores (scoring each emotion) have been used in classifying radical speech with reasonable success, with single class achieving an F-score of 85% and multi-class achieving 87% [5].

It is important to note that the best classifiers (F-Score 92%) used other features along with sentiment [5, 10]. These features include semantic patterns and semantic network detection. Semantic patterns are found through word embedding models that can look at how words are associated with each other, and semantic patterns are found through making semantic word graphs that show how words are related. These features make good sense as they capture more information about the language used, what words are typically used together, and how they are related. If time allows, it would be interesting to see if we could capture this information and take a more comprehensive look at radical speech.

3.3 Existing Graph-Based Approaches

In using NLP to study online hate, graphs have already been used in a few cases [8, 10, 18, 20]. The typical use case is creating semantic graphs for language

study. These graphs can help compute and visualize the relationship between different words, which is especially helpful when studying niche extremist speech that will not behave in the same way as regular speech [8, 10, 18, 20]. A single paper used graphs in another way to check the centrality of forum posters to do hotspot detection [4]. They paired graphs with clustering to find the most active and interconnected users.

For our purposes, we want to use graphs to represent online forum threads/users and the relationships between them. Graphs are particularly useful in this task because we can not only show which agents are connected but also by how much. Further, the graphical representation will allow us to leverage the tools in graph theory for our analysis.

There are a few examples of graphs being used in a way that is analogous to our work. When we look at the spread of sentiment, we will borrow some tools from epidemiological work. Since we want to see how sentiment spreads, our work may look similar to graph-based disease modeling, which has been used for some time but has seen a resurgence in the Covid-19 era [21, 22]. Graphs are especially useful in how disease spreads because we can model who is connected and by how much. The graph-based methods can consider things like social distancing, masking, and vaccination rate better than statistical models because they focus on the connections between individual agents [22].

Specifically, graphs have been used extensively in Social Network Analysis (SNA) [23]. SNA is a method of analyzing social networks through the joint use of networks and graph theory. SNA can be used to look at community clustering, information diffusion, identifying influential spreaders, and finding central members in the network [23]. All of these applications will be useful to us. Our graphs can look at community clustering and network modularity to identify subcommunities or echo chambers. The spread of sentiment may be significantly linked to information diffusion and may also behave in a very similar fashion. Identifying influential spreaders and other central members will be critical in modeling sentiment spread; we have already seen how the original poster dramatically influences the tone of the thread in previous work.

3.4 Centrality Metrics

Abdul Majeed et al. give a detailed breakdown of the various centrality metrics used in Social Network Analysis and what they mean [23]. This is an excellent place to start analyzing our graphs because these metrics work with information diffusion, which may be closely linked to sentiment diffusion. The breakdown of each metric described by Abdul Majeed and Ibtisam Rauf will be given below.

Closeness centrality (CC) generally measures how quickly a user or entity can access a large number of entities in a network. CC value can measure information spread or diffusion from a node. An entity with a high CC generally has four characteristics:

1. It has quick access to other entities in a network.
2. It has a short path to reach other entities.
3. It is close to other entities in a network.
4. It has high visibility about what is happening in the network.

Degree centrality (DC) represents the number of direct relationships of an entity in a network. A node with a high degree centrality has six properties:

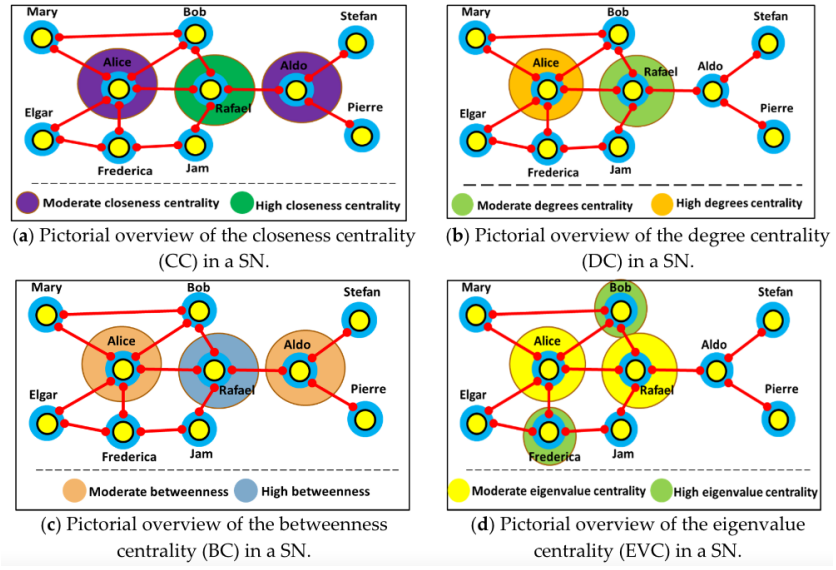
1. It is regarded as an active user in a network.
2. It often performs the role of a connector or hub in a network.
3. It is not generally the most connected entity in a network (an entity may have a substantial relationship, most of which refers to low-level entities).
4. It might be in a privileged position in a network.
5. It may have alternative paths to satisfy organizational requirements and depend less on other individuals in a network.
6. It can often be regarded as third parties or deal makers.

Betweenness centrality (BC) mainly identifies a user's placement within a graph in terms of its capacity to connect to other users or users' groups in a graph/network. An entity with a high BC generally has three characteristics:

1. It holds a favored or powerful position in a network.
2. It is prone to a single point of failure, i.e., take the single betweenness spanner out of a network, and you sever ties between cliques.
3. It has a significant amount of influence over what happens in a network.

Eigenvalue centrality (EVC) measures how close a user is to other highly close entities within a network. A high EVC has two main properties:

1. It represents an actor more central to the main pattern of distances among all entities in a network.
2. It is an appropriate measure of one aspect of centrality in terms of positional advantages.



(**a**) Pictorial overview of the closeness centrality (CC) in a SN.

(**b**) Pictorial overview of the degree centrality (DC) in a SN.

(**c**) Pictorial overview of the betweenness centrality (BC) in a SN.

(**d**) Pictorial overview of the eigenvalue centrality (EVC) in a SN.

Figure 4: Visualization of Centrality Metrics (*reproduced from* [23])

Centrality metrics used in other work [24–26] but not discussed by Abdul Majeed et al. is PageRank centrality (PRC) and Information centrality(IC). PRC is similar to Eigenvalue centrality in that it goes beyond looking at immediate links to a node, but it includes link direction as well [25]. PRC is essentially EVC for directed graphs and will be an indication of import/influential nodes.

Information centrality looks at which nodes are most influential to the flow of information through a network. IC combines eigenvalue centrality and closeness centrality by looking at all the paths that originate from a node and then weighting them inversely to their length [26]. Essentially, it is a measure of connectivity and closeness.

3.5 Echo Chambers and Othering of Individuals

An echo chamber inside of a social network will be a highly connected and tightly clustered collection of nodes that will often form a subgraph due lack of connection to members outside the echo chamber [23, 24]. They can be found using centrality and clustering metrics to look at network modularity [24]

Garimella et al.'s work [24] on studying political echo chambers on Twitter uses PageRank centrality and clustering coefficient to identify echo chambers. These metrics allow them to find tightly clustered, highly connected collections of nodes which indicates an echo chamber. They then verify they have found an echo chamber by looking at the political polarity of the nodes within it to make sure it is homogeneous in political leaning.

They further look into gatekeepers inside their social network as well. They identify a gatekeeper as a node that receives mixed polarity messages from other nodes but only responds with singular polarity back. They work as a filter that can take in messages of any political leaning but can only give back messages of one. This behavior is uncompromisingly exclusive to individuals outside of the gatekeeper's community and can keep people of opposing viewpoints out. They find that gatekeepers have high connectivity (high PageRank) but lost clustering coefficient, meaning they are connected to many other nodes but are

not embedded in any singular community [24].

3.6 Social Contagion Models

Looking at the spread of information inside of a network has been researched before; primarily by advertisers interested in seeing how to spread information about their products [7,8,23,27–29]. These studies focus on the term information diffusion, which is how information is spread from member to member or place to place through interactions. Information diffusion is looked at through its three main elements: information sender, information receiver, and medium of diffusion [27,28]. Looking at the behavior of senders and receivers in the context of their medium allows for the modeling of information diffusion.

One of the simplest models of information diffusion is the Two-Step Flow Model [27]. In the first step of diffusion, highly connected influential members of a community adopt an idea. In the second step, the members closest/most influential by these members adopt the idea they are using. An example of this model is given in Al-Taie's "Information Diffusion in Social Networks." chapter 8 [27], an example of this model is also given in figure 5:

"Shop owners who are central in their local networks with the highest in-degree scores... talk to their friends, customers, and social connections about the software and why it is a new trend in the market. Their immediate neighbors, in turn, will talk about the product to their neighbors and so on until finally the news about the product is spread to a large population of users in the network."

The next model, also described by Al-Taie [27], fits very well into the existing Graph-Based Approach to Studying the Spread of Radical Online Sentiment we have built. The Social Contagion Model treats the spread of ideas like a disease, exactly as we intend to do with sentiment [27]. It starts off with a small group of people having the idea and does a chain-like spreading to the people closest to them that grows outward, "infecting" more and more people.

The Social Contagion Model is more advanced than the Two-Step and takes

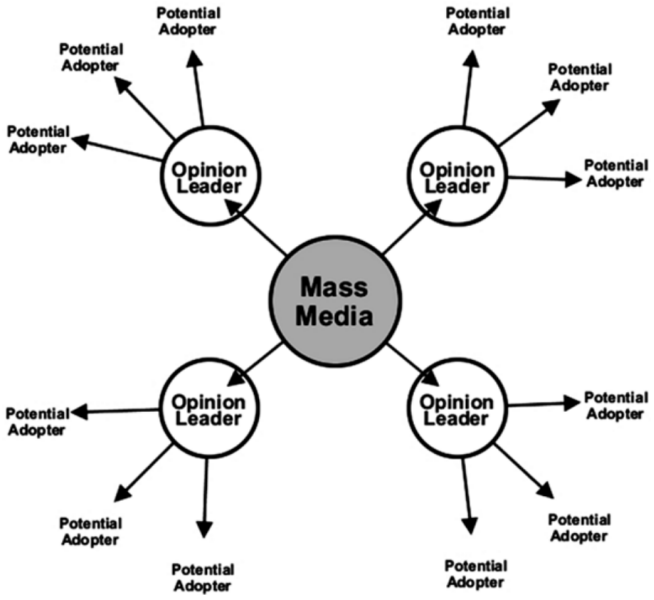


Figure 5: Two-Step Flow Model [27]

into account the adoption rate, types of adopters, adoption thresholds, amount of exposure, and idea momentum. The Social Contagion Model will produce a sigmoidal S-curve very similar to the SIRS model [21,22,27] where early adopters will spread the idea very quickly until it hits a critical mass to go mainstream and get constant linear growth until it has reached everyone (or everyone who will accept the idea) and then tapers off, seen in figure 6.

3.7 Defining Radical Speech

To first talk about what is radical speech, we need first to define what is radical or what makes something radical. Torregrosa et al. [6] gives a very descriptive definition of radicalization and extremism in the context of NLP research done on extremism. He notes that "there is no academic consensus about the definitions of extremism and radicalization," and the terms are used synonymously in most work, but he gives concrete and separate definitions for both words.

Torregrosa states: "*Radicalization* was born during the 18th century as a

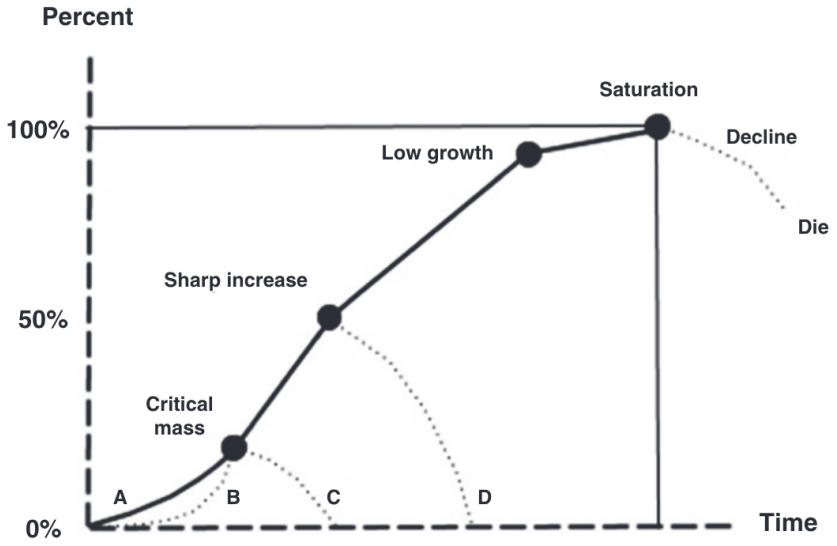


Figure 6: Social Contagion Curve Example [27]

way to define a movement against the establishment, but not inherently violent or positioned against democratic values. Meanwhile, the concept of *extremism* refers to an anti-democratic movement and stands against "all those who do not embrace its dogmatic recipe for a transformation of society." [6]

Using our words carefully now, we can say that radical speech is speech that goes against the current establishment. It is not inherently bad, it is simply contrarian. On the other hand, extremist speech (which can be radical speech) will adhere to a transformative ideology that will not accept those who do accept it. Given these two terms and their definitions, we should retool our terminology to use *extremism* and *extremist speech* rather than radicalization and radical speech. What we are looking for is this harmful extremist speech rather than radical speech, which may be an indicator of extremism but is not harmful in itself (Prompting universal healthcare and free education could be considered radical speech in the U.S.).

3.8 Stance Detection

A 2021 literature survey by Küçük et al. gives an extensive overview of the cutting-edge work done in stance detection [30]. Küçük et al. define stance detection as identifying the relationship between a target pair of texts into the categories of favor, against, neither, and sometimes neutral. The neutral category is used when there is not enough information in the text to identify any stance taken. This is categorically different from neither, which indicates the text does not take a stance for or against (analogous to zero vs. null).

Stance detection itself can be broken down into multiple categories identified by Küçük et al. [30, 31]. These categories are:

1. Stance detection: Determining the stance relationship between a target pair of texts.

2. Multi-target stance detection: Determining the stance relationship between a piece of text and multiple targets.

3. Cross-target stance detection: Classifying the stance of a piece of text by comparing it to other pieces of text with a known stance.

4. Rumor stance classification: Determining the stance of a piece of text towards a rumor or a rumored pair. The piece of text can be categorized as Supporting, Denying, Querying, and Commenting

5. Fake news stance detection: Comparing the headline of a news article to the body text of a different article on the same topic determining whether other articles on the same topic agree with the statement made in the headline. The body of text being compared to the headlines can be categorized into Agrees, Disagrees, Discusses (the same topic), and Unrelated.

These stance detection techniques are used in many areas, such as opinion/survey polling, market trend analysis/forecasting, recommendation systems, rumor classification, and fake news detection.

To truly understand the use of stance detection, we need to know how it is different from sentiment analysis. Though sentiment is a feature used in stance detection, stance detection goes beyond finding the emotional polarity of text [30, 31]. This can be seen in the cases of two comments having the same negative sentiment because they are disagreeing with each other and two comments having opposite sentiments because one is a comment in agreement (with positive sentiment) with a negative comment. Raw sentiment analysis will tell us if these comments are in agreement or not, but stance detection will.

Finally, Küçük et al. go over what methods are used for stance detection. Almost all studies view stance detection as a classification problem and use machine learning, deep learning, or ensemble methods [30,31]. The features used in these classifiers include lexical features (bag-of-words, n-grams), interaction features (likes, dislikes, replies, retweets), sentiment features, word embeddings (word2vec), topic modeling, and part of speech tags. Küçük et al identify the most common models used in the papers they reviewed in a word cloud. As we can see in figure 7, machine learning models are most commonly used with deep learning networks and ensemble methods being used to a lesser degree.

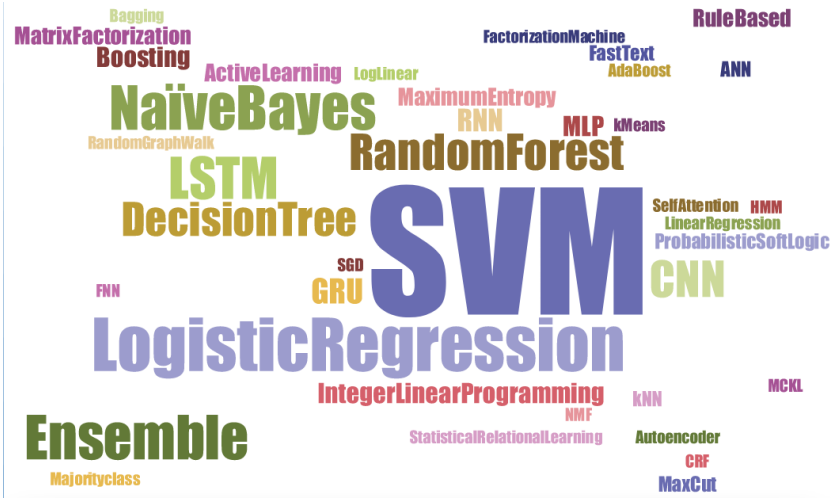


Figure 7: Stance Detection Model Frequency [30]

3.9 RoBERTa Model

The Robustly optimized BERT approach (RoBERTa) model is a retraining of Google's BERT large language model that gives it state-of-the-art performance on NLP tasks [32–34]. Key differences between the original BERT training and RoBERTa are the amount of data trained on, batch size, and training input. RoBERTA was trained on a large number of corpora totaling to 10 times the amount of tests BERT was trained on. The larger amount of data allowed for a large batch size as well, with BERT training on 256 samples at a time and RoBERTa training on 2,000.

The larger data set and batch size improved performance, but RoBERTa goes a step further and also changes the model's input [33]. First, RoBERTa trains on full sentences rather than sentence fragments like BERT. This allows RoBERTa to see more context and better understand the relationship between words. Similarly, BERT's word embedding space has a vocabulary size of 30,000 compared to RoBERTa's 50,000. Lastly, RoBERTa implemented a dynamic masking scheme on all of its inputs; masking different tokens in the input embedding instead of masking the same ones each time (like BERT).

All of these training changes allowed RoBERTa to outperform BERT in key language model metrics such as the General Language Understanding Evaluation (GLUE) benchmark, the Stanford Question Answering Dataset (SQuAD), and the ReAding Comprehension from Examinations (RACE) giving it state of the art competitive performance to other large language models [33]. RoBERTa has even been found to outperform domain-specific language models. A study done by Ankur Sinha et al. found that a pre-trained RoBERTa model could outperform finBERT, BERT trained on financial news when it came to sentiment analysis on text relating to finance [34].

4 Proposed Approach

Our approach studies the diffusion of sentiment and online community dynamics using social contagion modeling and social network analysis. These methods have been thoroughly developed analytically and proven to be effective experimentally on both real and simulated graphs [35].

We start by constructing graphs using data from AZSecure dark web forums [11]. These graphs are constructed on inter- and intra-thread levels where, on the inter-thread level, entire threads are nodes connected to each other by edges weighted by the number of shared members (figure 8). On the intra-thread level, nodes are members of a thread, and they are connected by edges weighted by how many replies are made to other members (figure 8).

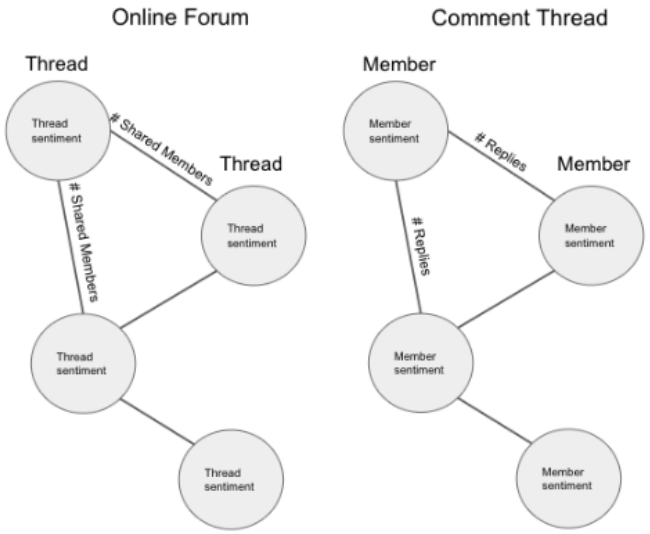


Figure 8: Visual Example of Inter-Thread (left) and Intra-Thread (right) Graph

Leveraging social contagion and social network analysis techniques, we study how sentiment spreads from member to member and graph to graph. We obtain the sentiment of every comment and every member in our data with the RoBERTa Large Language Model [33] and observe the changes in sentiment with the evolution of our graphs. We also identify various community dynamics, such

as the most influential spreaders using network centrality.

After we understand the behavior of our graphs and how sentiment spreads within them, we can make a model to classify the sentiment of a node within a graph and predict the sentiment of the next member to join the graph. We do this by using the most important features we found to determine the sentiment of a comment or member to train a classification algorithm. This algorithm will then be tested against our real data to find its performance.

4.1 Example Graph

To further illustrate our method, we provide a sample graph created from our data in figure 9. We create an intra-thread graph from a chain of 107 comments from 32 members of the forum. Our analysis includes a graph for all 1,973 comment threads from the online forum. This suggests a relatively complex conversation with many different participants involved. By analyzing the intra-thread graph, we aim to identify patterns and relationships that were not immediately apparent from simply reading the comments. To visualize a group of members with similar sentiments, each member (node) of the group is color coded and labeled based on the sentiment of their last comment. This is useful for understanding a group's overall sentiment or mood and identifying members who may be particularly positive or negative in their comments. In figure 9, nodes are color coded using a green-yellow-red color scale, with green indicating positive sentiment, yellow indicating neutral sentiment, and red indicating negative sentiment. Directed edges show the direction of the communication between members, which other members replied to each other. The weights of the edges have been recorded but could not be shown to minimize visual clutter.

4.2 Experiments

RQ1: To test the extent members are connected to other members and threads are connected to other threads with a similar sentiment, we create graphs on both intra-thread and inter-thread levels. Each node had its member/thread

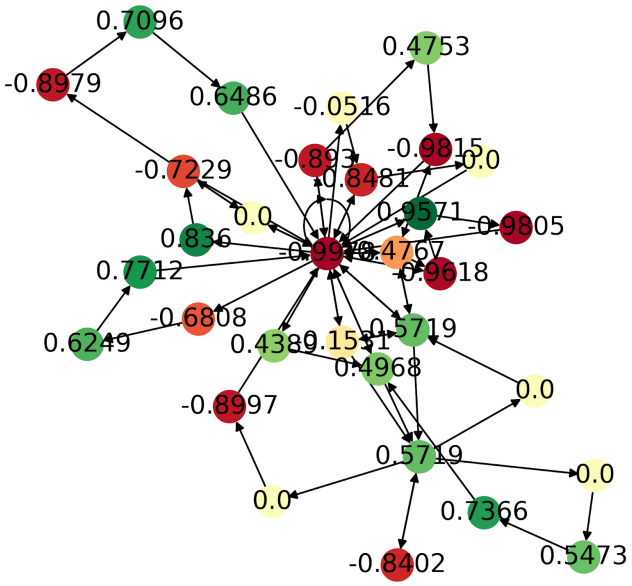


Figure 9: Example Intra-Thread Graph

sentiment stored and we observed all the connections between nodes (edges). We tallied our results to see if connections have a tendency to start and end at similar sentiments (positive to positive or negative to negative).

RQ2: To observe the most influential members in the thread and how they shape the overall sentiment of their thread, we first looked at the information centrality of each member in our constructed graph. We then compared the sentiment of the central member or members to the overall sentiment of the graph to measure their influence.

RQ3: To understand the behavior of our threads over time, we performed temporal analysis on our intra-thread graphs. We evolved the graphs by following the timestamps in our data; when a new member joined and when a new comment was made. The sentiment of the graphs and members in them was recorded as more members joined and commented. Finally, we developed summary statistics on how sentiment changes as well as what interactions occur

as the graphs evolve.

RQ4: To determine if the behavior of our graphs could be modeled, we created a model that can classify the sentiment of a node as positive or negative based on features from the nodes connected to it. Then we had the classifier act as a contagion model and make predictions on new nodes joining the graph. Finally, we compared the model results to our real data to calculate its overall performance.

5 Experimental Evaluation and Results

5.1 Data Set

To address our research questions, we utilize the AZSecure collection of dark web forums [11], specifically the *Gawaher* data set (see table 1), which is an "An English language Islamic forum" dedicated to discussions made available for open-source research.

Data Pre-processing: We set a minimum threshold for the number of comments and active members to ensure that the data set contains only meaningful and well-populated threads. Therefore, we include threads with 30 responses and at least two active members (members who make responses to other members). This step resulted in a drastic reduction of the data set from approximately 50,000 to 2,000 threads which contain 137,000 comments in total. Figure 10 gives the distribution of thread length, which throws light on the size and structure of the threads in our data set. Most threads are short, with a median thread length of two, and 20,540 threads have only a single comment.

We analyze the distribution of the mean sentiment in our data to study the class imbalance (figure 11). This information can help us understand the nature of the data we are working with to determine if it is appropriate for our study. The histogram shows that the sample is skewed toward negative sentiment with an overall mean sentiment of -0.164, suggesting that more threads contain extremist and generally negative messages. Approximately 80% of the threads

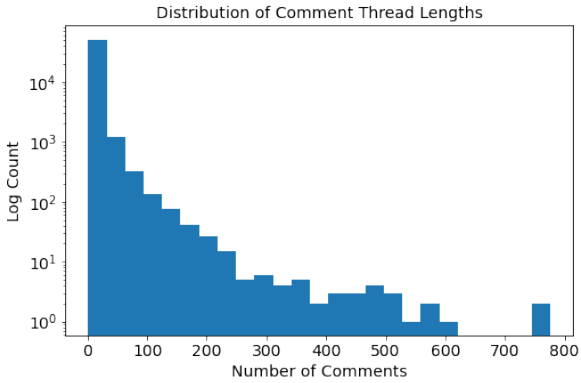


Figure 10: Distribution of Comment Thread Lengths

show negative mean sentiment, whereas the other 20% have a positive one.

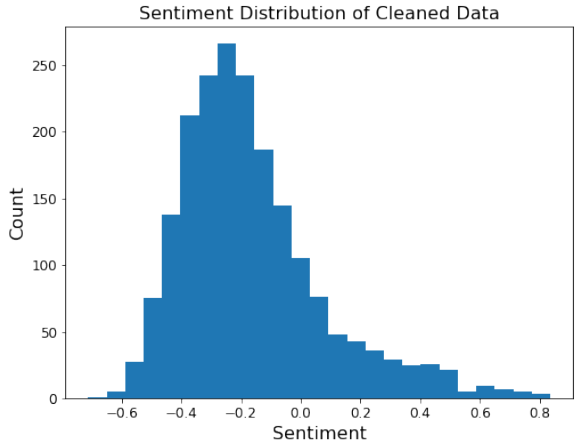


Figure 11: Distribution of Mean Sentiment of All Threads in Cleaned Data

We are aware of the class imbalance and note that it does not affect our current study as we only examine if members of the same class tend to connect. However, with a larger data set, we plan to perform a deeper analysis and apply over-sampling or under-sampling techniques to address the class imbalance problem.

MsgID	ThreadID	ThreadName	MemID	Message	Year	Month	Day	Time	PrevThread MsgID
100013	2583	Islam in Turkey	elif74	yes, this is very sad. but there are more people converted to Islam than christianity in turkey. this is better idea.... the problem with their families, they dont teach anything to their children and just having baby to have. i mean w/ no responsibility. the result is this.	2005	3	26	59:00.0	24378
1000177	533029	Just When You Thought Iran Couldn't Get Any Crazier	troof	¡quote¿Yasnov, on Oct 26 2008, 02:13 AM, said: it does not take a rocket science to figure it out. ask any 5 year old kid, they will also give the same answer: israel future is bleak. are you against the idea of introducing democracy to israel? Umm, what does that have to do with your strange claim that the world will want and permit israel to be wiped off the map?	2008	10	26	04:00.0	993099
1000246	533029	Just When You Thought Iran Couldn't Get Any Crazier	Yasnov	¡quote¿troof, on Oct 26 2008, 06:04 PM, said: Umm, what does that have to do with your strange claim that the world will want and permit israel to be wiped off the map?what made you think that the world will never wake up? wassalam, y	2008	10	26	08:00.0	993099
100026	10096	What Would You Do, If Someone (opposite)	Teakster	Salaam, I would jump in the Teakster mobile and run the evil monkies over!!! Wasalaam	2005	3	26	16:00.0	94941
100028	9186	Whats Your Warning Level?	Teakster	Salaam guys, I sometimes hate the mods...! You try to post a good topic or have fun..... then they come waving the mod badges everywhere and close everything down! Man.....Why.....?!?!?!¬†¬† I'm gonna get a warning for this, aren't I?	2005	3	26	25:00.0	87806

Table 1: Snapshot of the Gawaher dataset extracted from AZsecure [11]. The dataset has from some forum members who sympathize with radical Islamic groups. Postings are organized into threads which generally indicate the topic under discussion. Each posting includes detailed metadata such as date, member name, message posted, thread ID, and member ID.

5.2 Results

5.2.1 RQ1

Intra-Thread Connections

This analysis is broken down into three categories: positive-positive connections, negative-negative connections, and positive-negative connections (figure 12). In addition, the analysis is split between unweighted and weighted connections. Unweighted connections check if a connection exists between two nodes while disregarding the degree of connection. Weighted connections, on the other hand, take into account the number of times members reply to one another, thus giving more weight to connections that occur more frequently.

Based on the results, it appears that most connections are between members of a similar sentiment, using both weighted and unweighted connection schemes. Likewise, weighting the connections increases the percentage of similar sentiment connections, suggesting that members with similar sentiments tend to reply to each other more frequently than those who do not. This analysis can be valuable for understanding the communication patterns and sentiments of a group, particularly within the context of specific comment threads.

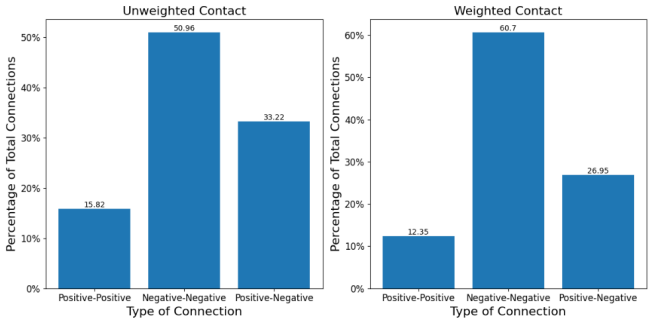


Figure 12: Distribution Weighted and Unweighted Connection Types in Intra-Thread Graph

Inter-Thread Connections

In this case, we examine the total number of connections between members

with similar sentiments across different comment threads. Based on our results, there is a similar pattern of behavior as in the intra-thread connections analysis, with the majority of connections being between members of a similar sentiment (figure 13). However, the percentage of mixed connections (positive-negative) is higher in the inter-thread analysis than in the intra-thread analysis. In addition, the analysis of weighted connections only shows a significant increase in positive-negative connections between threads (3%), which comes from a decrease in negative-negative connections. This suggests that threads with negative sentiments tend to share more members than threads with positive or different sentiments.

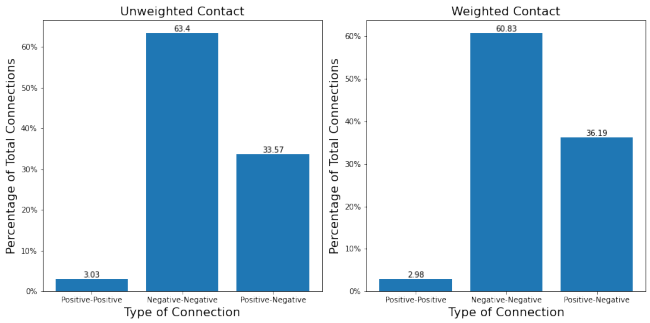


Figure 13: Distribution Weighted and Unweighted Connection Types in Inter-Thread Graph

5.2.2 RQ2

Centrality

We use a graph-based approach to identify the most central or well-connected nodes in the network graph. Here, we assert that the most central member of a given graph will have the most influence over the general sentiment. To find central members, we use information centrality. Information centrality is a measure of a node's importance in a network based on how much information flows through it. In our graph, instead of information, sentiment flows between nodes. Therefore, nodes with high information centrality are those that are most

likely to receive and disseminate sentiment in the network. This information can be valuable for understanding the dynamics of the group or community and for identifying potential leaders or influencers who may be able to shape the sentiment of the group.

We analyze the influence of every commenter on the sentiment of the thread using information centrality for network graphs formed on the intra-thread level (figure 14). Additionally, we looked at the most active member in each by observing who made the most comments (figure 15). Results show that the original poster (0th commenter) is the central node and most active member in the majority of graphs on the intra-thread level, indicating that they had the most influence over the overall sentiment of the thread. This suggests that new commenters tended to follow the initial topic or sentiment set by the original poster.

We also looked at the distribution of the difference between the original poster's sentiment and the mean sentiment of the thread in figure 16. This analysis shows that the average sentiment of the thread generally matches the sentiment of the original poster. This further supports the idea that the original poster has the most influence over the sentiment of the thread and that new commenters tend to follow the initial topic or sentiment set by the original poster.

Our analysis of commenter centrality at the intra-thread level sheds light on why there is less shared sentiment on the inter-thread level. Members tend to engage with the topic inside a given thread, and even if they switch between threads, they will switch to the other thread's topic and thus switch to the thread's sentiment. This indicates that members are influenced by the specific topic and sentiment of each thread and that the sentiment is not necessarily carried over from the previous thread they were in.

These insights can help us understand how sentiment is influenced and changes over time in a community or group. They also highlight the importance of the topic and sentiment of each thread in shaping the overall sentiment of the community or group.

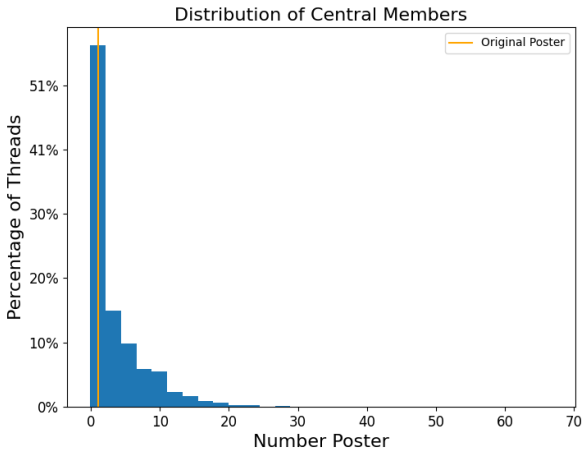


Figure 14: Distribution of Central Members in All Intra-Thread Graphs. Number Poster indicates the order of the posters

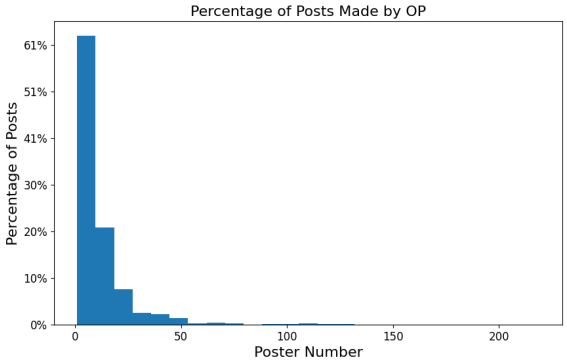


Figure 15: Distribution of Most Active Commenters

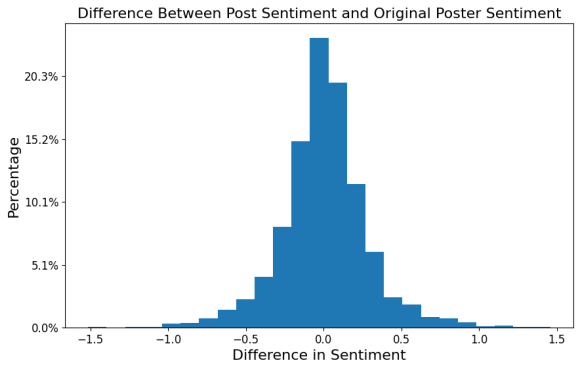


Figure 16: Distribution of the Difference in Sentiment Between Original Poster and Thread Mean Sentiment

Central Member	Original Poster		Not Original Poster	
Statistic	Mean	Std	Mean	Std
Thread Length	63.88	56.48	72.65	69.65
Number of Comments	5.15	6.15	3.85	2.63
Number of OP Comments	5.68	10.67	3.57	5.08

Table 2: Statistics on Threads with Central Original Poster and Non-Central Original Poster

Centrality for Non-Original Posters

The original poster being the most central member of a thread is an intuitive finding. We performed a deeper analysis of threads where the original poster was not the central poster. From our findings in table 2, we discovered that threads that had a non-original poster as their central member tended to be longer threads, with each member (including the original poster) only contributing a few comments. This suggests that if the original poster is not the central member, then there is no central member of the thread. The thread is long and the number of comments per member is few, so centrality changes between members and does not settle on a single one.

5.2.3 RQ3

Temporal Analysis

To study the temporal nature of our threads, we measured the mean sentiment, sentiment standard deviation, and central member every time step. We used thread interactions as our time steps. That is, every time a comment is added to the thread, whether it be a new member joining and making a comment or existing members replying to each other, we record that as a step. We can see some example time series in figure 17.

For both the mean sentiment and sentiment standard deviation, we saw similar behavior. The time series started out chaotic as there were few interactions in the thread; every interaction had a large sway on the overall sentiment. As time went on, the mean and standard deviation regressed to a mean value and varied little from it. As we can see in the aforementioned figure, the long threads with more interactions are more consistent at a value they regressed to, while shorter threads with few interactions are chaotic.

These findings track well with our previous observations in **RQ2**, where we saw the overall sentiment of the threads tended to be distributed around the sentiment of the original poster. The original poster of the thread sets the sentiment, and the rest of the posters tend to follow suit.

Further, we broke our analysis down by original posters with positive sentiment and negative sentiment. We wanted to discern if there was different behavior in the time series that started off negative rather than positive; having the thread started off as radical. Our first hypothesis was that threads with negative sentiment would have more discourse in them and have less variation in sentiment, as everyone would have a tendency also to have a negative sentiment. This ended up being the case with the mean standard deviation of threads with a negative original poster being 0.50 and 0.59 for threads with a positive original poster. The different distributions can be seen in figure 18.

We also computed the difference in mean starting and ending sentiment of our threads to see if threads with a positive original poster were more likely to

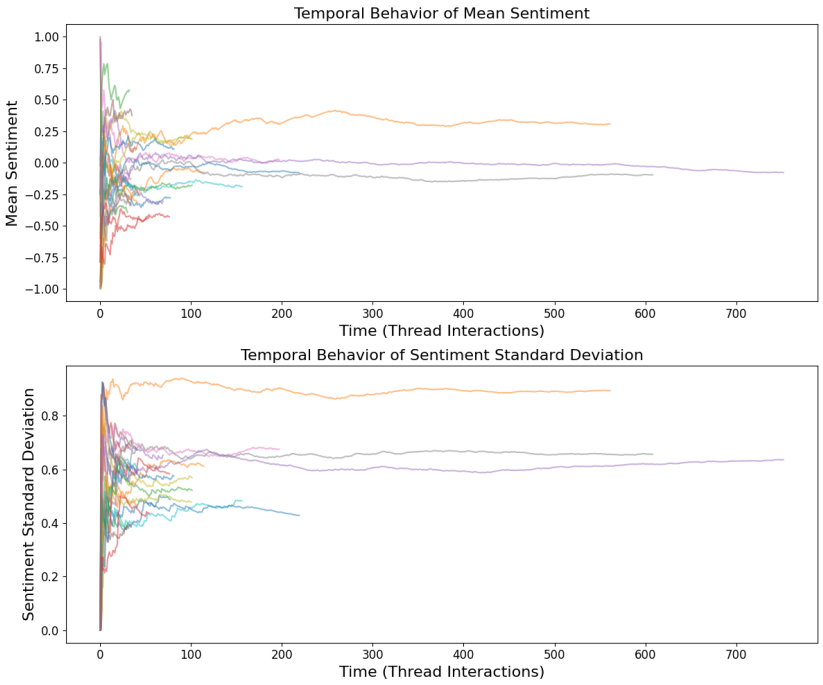


Figure 17: 25 Example Time Series for Temporal Analysis

end with negative sentiment. It might be the case that even if the post starts out with positive sentiment, disagreement, and argumentation can pull it down to having a negative sentiment. Our results can be seen in figure 19 where the distribution of starting and ending sentiment difference is centered close to zero for threads that start negative, while threads with a positive original poster had a difference centered around -0.5. This supports our assumption that threads that start negative tend to stay negative more than threads that start positive tend to stay positive.

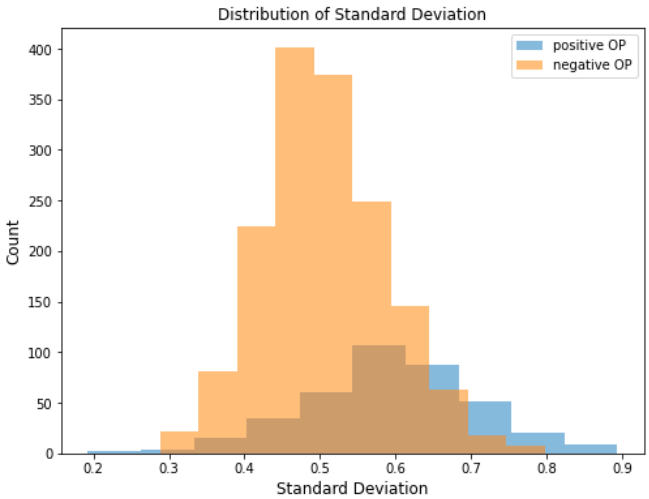


Figure 18: Distribution of Sentiment Standard Deviation

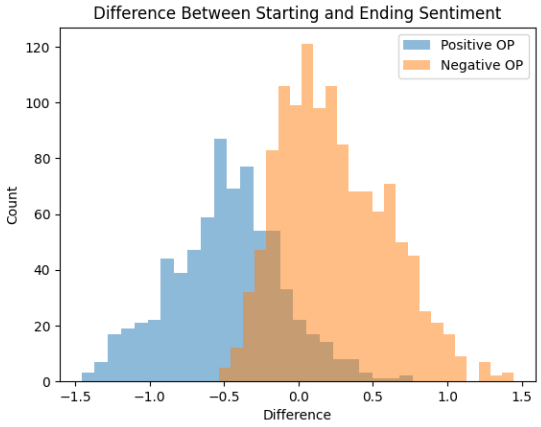


Figure 19: Distribution of Starting and Ending Sentiment

Lastly, we studied the most central member of our threads over time. Our findings support what we discovered in **RQ2**, with the original poster dominating as the most central member for the majority of time steps (figure 20).

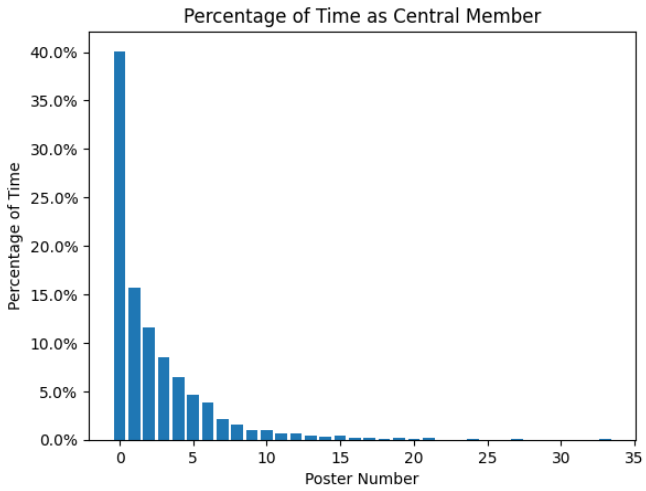


Figure 20: Percentage of Time Steps as Central Member by Poster Number

5.3 RQ4

Classification

With an understanding of how our social graphs behaved, we could model their behavior. The first model we made was a classifier that could determine the sentiment of a node based on its connected nodes seen in figure 21. Determining the exact sentiment of a node was intractable and that level of detail was unnecessary for our study. Thus, we kept it as a binary classification problem where we classified a node as having positive or negative (1 or -1) sentiment based on the nodes it was connected to.

We used a random forest classifier for our model that used features of connected node sentiments, edge weights, and centrality of both the target node and the nodes connected to it. To test the validity of our model, we also implemented a naive approach where we took the mode of the sentiments of the

	Naive (KNN)	Random Forest
F1 Score	0.54	0.84
Accuracy	0.58	0.87
TPR	0.60	0.81
FPR	0.43	0.08

Table 3: Naive vs. Random Forest Classification Performance

connected nodes, similar to the K nearest neighbors. As seen in figure 3, our random forest classifier outperformed the naive approach with an f1 score of 0.84 on our test set to the naive 0.54, which we argue is equivalent to random guessing.

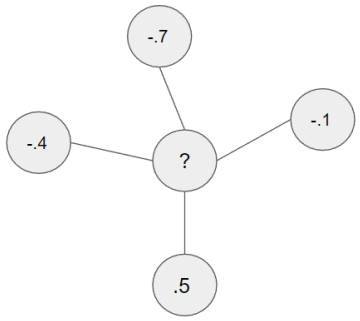


Figure 21: Example of Node Sentiment Classification Problem

Prediction (Contagion)

We used the same classification model to predict the sentiment of new commenters to the thread. Based on our understanding of contagion, we could classify the sentiment of a node coming into the graph by knowing the sentiment of the nodes currently existing in the graph as well as how the new node would connect to the existing ones (how a new member of a thread would comment to existing members). With temporal information on all of our graphs, we started with a set amount of seed nodes and had our classification model use them to classify the sentiment of the next node to come into the thread. We

had it continually classify nodes (Using the nodes it already classified as more information to classify the next one) until all nodes in the thread were classified. After it made predictions on all nodes, we determined the performance by comparing it to the actual node sentiments. An example of our problem can be seen in figure 22.

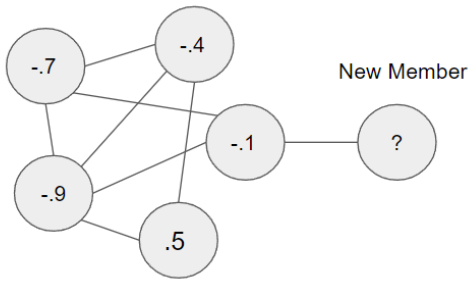


Figure 22: Sentiment Prediction Example

We ran our sentiment prediction/contagion model several times, seeding it with a different number of known nodes to see how it would change performance. The results of our model can be seen in figures 23, 24, and 25 as a distribution of model accuracy for each thread. The final statistics for the model can be seen in table 4, where we see the f1 score improve with a greater number of seed nodes.

	1 Starting Node	3 Starting Nodes	5 Starting Nodes
F1 Score	0.60	0.61	0.64
Accuracy	0.71	0.72	0.72
TPR	0.61	0.62	0.64
FPR	0.22	0.23	0.22

Table 4: Performance Statistics of Model Based on Number of Starting Nodes

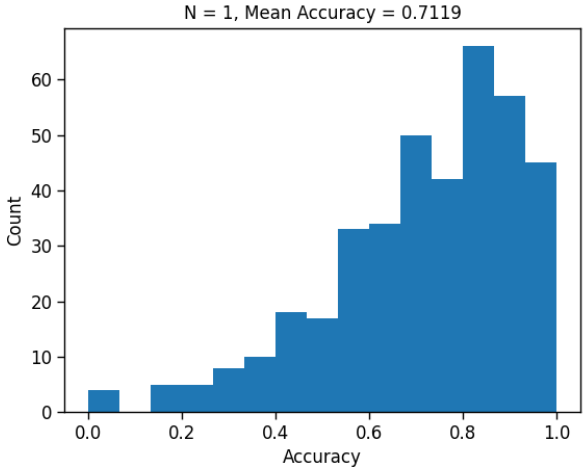


Figure 23: Accuracy Distribution of Model for N = 1 Seed Nodes

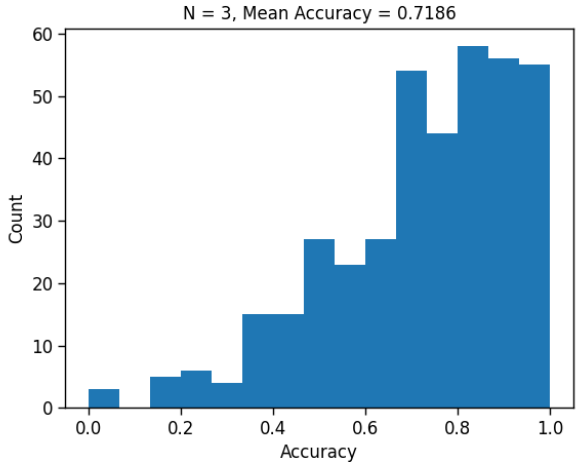


Figure 24: Accuracy Distribution of Model for N = 3 Seed Nodes

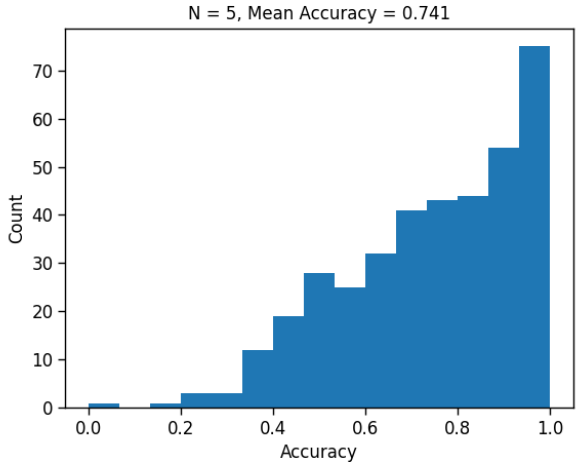


Figure 25: Accuracy Distribution of Model for N = 5 Seed Nodes

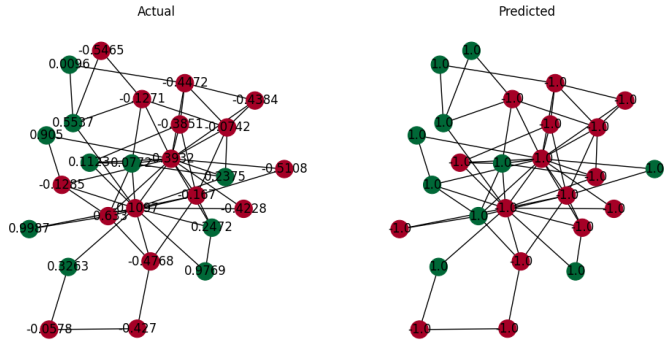


Figure 26: Example of Full Predicted Graph (Starting with N = 5 Seed Nodes) Compared to Actual Graph

5.4 Extended Analysis

Inter-Thread Clustering

While conducting research for this thesis, we had additional questions and explored additional areas outside of our research questions. One of these areas was looking into community formation and clustering on the inter-thread level. Our hypothesis was that we would find groups of posters that would visit the same threads. We could find these groups by clustering all posters by the threads they visited and we would find groups that had similar sentiments and visited similar topics.

This work ended up being inconclusive. We performed clustering of each member by threads visited and found there was no clear clustering of posters seen in figure 27. This was a curious result so we did more analysis and found that the median number of threads posters interacted with was 3 (figure 28). If the average poster is only interacting with 3 threads out of the 1,973 in our data set, it is unlikely that we would find posters consistently visiting the same threads.

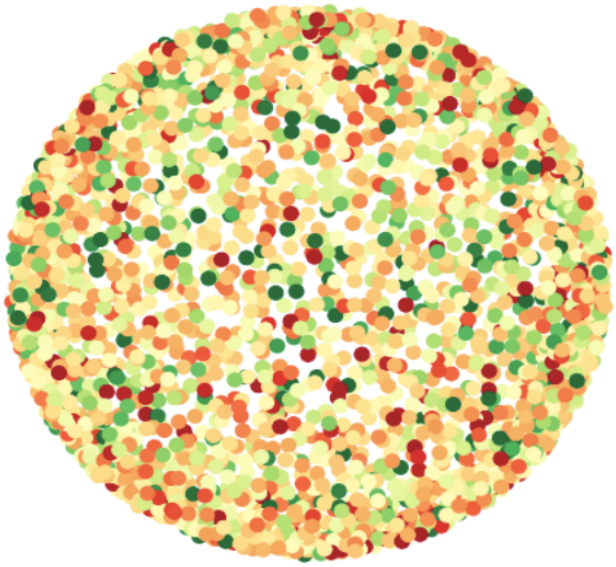


Figure 27: Clustering of all Posters by Shared Threads, Color Coded by Sentiment.

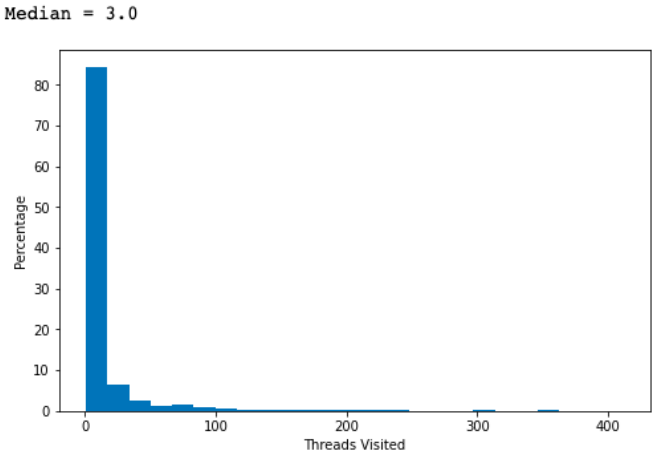


Figure 28: Distribution of Number of Threads Visited by Each Poster

Simulation Work

Another avenue of work that did not become a research question was simulation work to create synthetic social graphs. The simulation consists of functions that can add members and replies to the graph as well as propagate the sentiment from one member to another. These simulations allowed us to create more graph structures and study more behavior than what we saw in the data. Unfortunately, due to time constraints, we could not fully flesh out this body of work but we do have preliminary results from one of our simulation experiments.

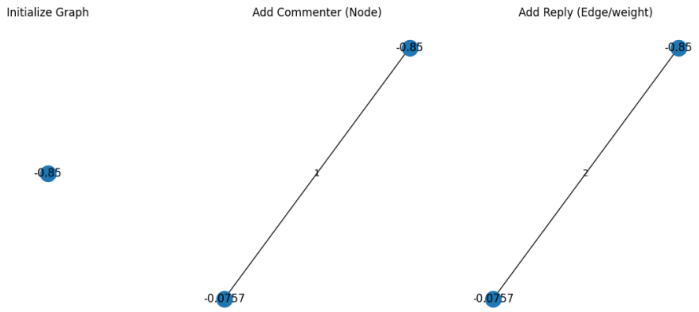


Figure 29: Simulation Functions

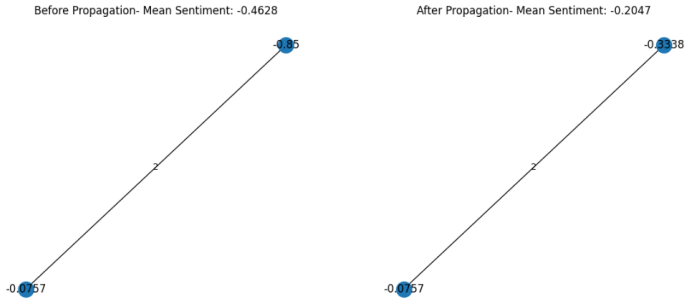


Figure 30: Propagation of Sentiment in Simulation

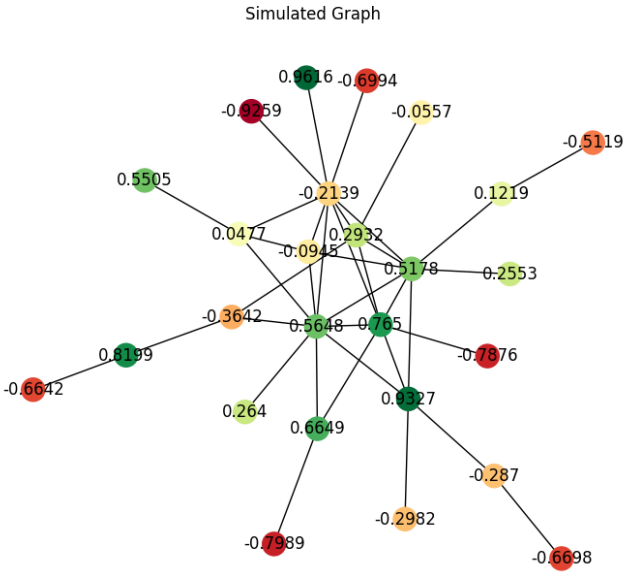


Figure 31: Simulated Graph Example

Simulated Graph Experiment

One of the experiments we performed with our simulation was the interactions of a tightly-knit toxic community with members outside of the said community. We started by creating a toxic community (figure 32) and then adding members to it and letting them make replies to the existing community (figure 33). We made the following assumptions about random members entering the toxic community: (i) members entering the community will have a random sentiment, (ii) sentiment will propagate through replies made to other members, (iii) even with a reply, members will have a 50% chance of not adopting the sentiment propagated to them. We allowed 30 new members to be added to the existing toxic community and let 60 new replies be made.

Similar to our time series analysis done in **RQ3**, we took the mean and standard deviation of the sentiment in our simulated graph every time step. Again we define time steps as the addition of another member (node) or reply (edge) to the graph. We saw similar results to our time series analysis in **RQ3** (figure 34). The time series is chaotic for the first few steps and then regresses

to a mean as time goes on. The toxic community started the mean sentiment off negative and it stayed negative as members joined and made replies. These initial results are promising from our simulation because they closely match the behavior of the actual data. We saw that the most central and well-connected member of a thread has the greatest influence on the sentiment of the thread over time, and the simulation repeats this behavior.

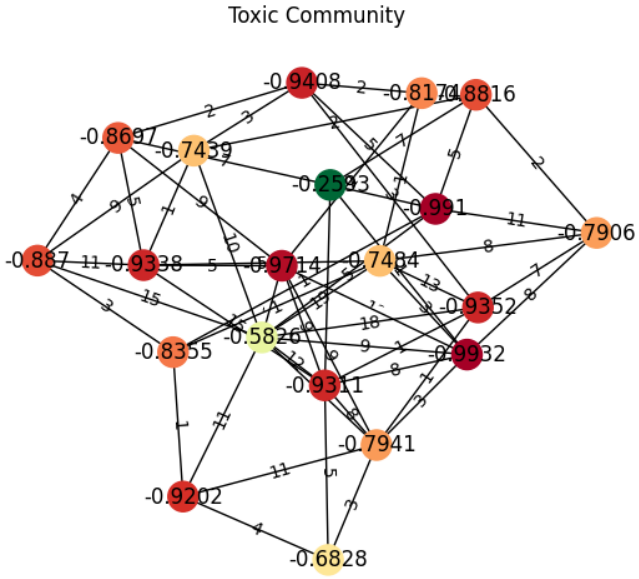


Figure 32: Simulated Toxic Community

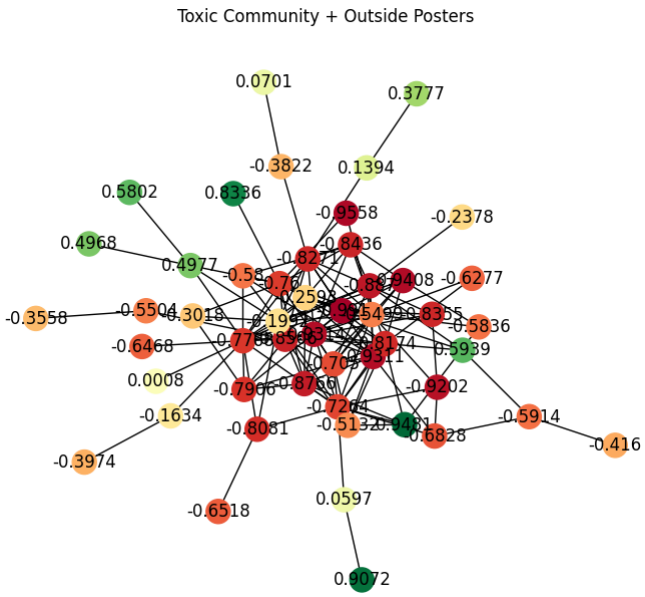


Figure 33: Simulated Toxic Community with Outside Members Added

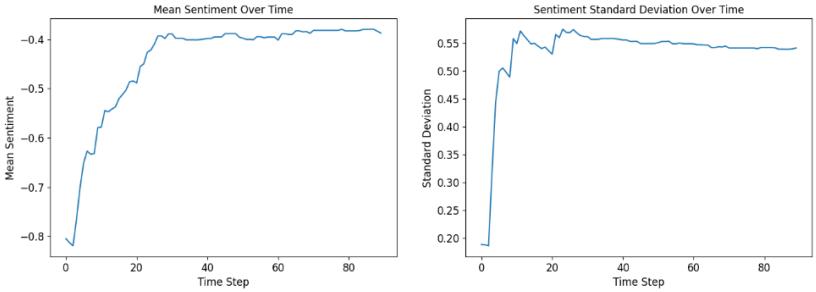


Figure 34: Time Series of Mean and Standard Deviation from Simulated Graph

6 Discussion, Analysis, and Conclusion

At the beginning of this thesis, we claimed that there is a gap in the current body of work on Online Hate Research. That is, the focus of the current work is only on studying language and ignores community dynamics. By answering all of our research questions, we have proven that looking at the community structure of online comment threads is vital to understanding them. By using graph theory to analyze the community structure of comment threads, we have determined how sentiment spreads between members of a thread, who the most influential members of a thread are, and have made classification and contagion models to determine the sentiment of members in a thread.

RQ1 proved our basic assumption that members within a thread and threads them self generally connected to other members/threads with similar sentiments. **RQ2** showed us that the most influential member of a thread could be determined using graph centrality, and the most central member is the original poster in the majority of cases. **RQ3** supported our findings in **RQ2** and allowed us to observe the temporal dynamics of these threads. All of our RQs lead into **RQ4**, where we used the information we discovered to model the sentiment in our graphs.

The performance of the models we created proves the significance of community features when studying online comment threads. We achieved comparable results to previously created models with a different feature set. The models we reviewed in the literature used linguistic and syntactic features such as word embeddings, part of speech tagging, and multiclass sentiment analysis [5, 10]. In this work, we used single-class sentiment analysis, centrality, and degree of connection. Our model was able to achieve an F1 of 0.84 with our network-based feature set to classify if a poster had positive or negative sentiment. This is equivalent to an aforementioned single-class sentiment model that achieved an F1 of 0.85 that used a set of NLP-based futures. Though more advanced models that use and can classify multi-class sentiment with an F1 of 0.92 are the cutting edge, we have proven network features are significant and should be

included in future models.

When using our classification model to predict sentiment, our contagion model performed worst with an F1 of 0.64. This makes good sense as there is much less information when a new node enters the graph compared to a node-set in an already existing one. Though this model needs refinement, we argue that its performance is still indicative of our graph-based features being significant in the prediction of sentiment and should be incorporated into future sentiment prediction/contagion models.

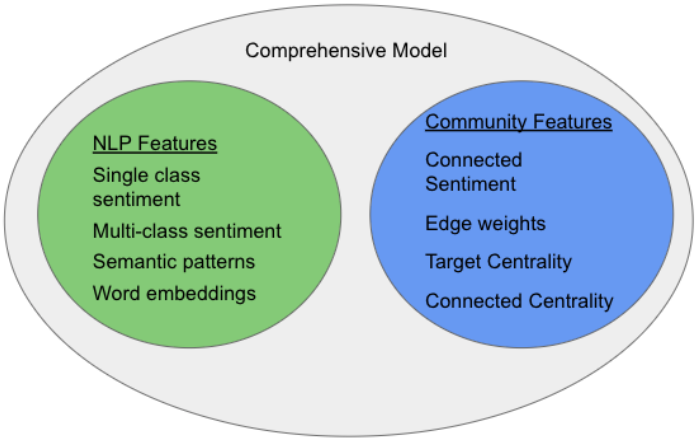


Figure 35: Proposed Future Model

6.1 Limitations

As previously mentioned, contextual sentiment analysis is an area of active research in natural language processing and our work will have the same faults as other work when it comes to the analysis of language with biased effect [5,6,10, 17,18]. When looking at the language used in our data set, we need to keep this in mind since we are looking at the language of a specific community. Extremist Jihadist language has its own vocabulary and connotes differently than general speech meaning any language model we use that is trained on general language will be error-prone. Since our work is based on studying sentiment, this will

impact our results. We need to account for where our tools work and don't work.

In a similar vein, contextual sentiment will impact the meaning of our mixed sentiment (negative to positive) connections. Having a comment with positive sentiment connected to a comment with negative sentiment could be a comment that agrees with or praises the negative comment. This would also be the propagation of the negative sentiment but with a mixed sentiment connection.

Something we can not account for is the spread of sentiment through lurkers. Lurkers are members of an online community that will read threads but not interact with them. Lurkers often make up a silent majority of online communities, with upwards of 90% of members reading threads but not interacting with them [36]. The influence the content of a thread has on lurkers is unknowable with our method because they do not exist in our data. Further, our analysis will miss all connections of lurkers in one thread becoming active members in another. They could carry the sentiment from the thread they read and never interacted with into a new thread and we could never know if the sentiment of that thread spread to another.

Finally, the scope of this work will be limited due to the size and specific nature of our data set. Since we are only at the proof of concept stage of this work, a small data set of known radical language was picked intentionally to develop our method before expanding to larger, more general data sets.

6.2 Future Work

We concluded our analysis after comparing positive and negative sentiment values. A more granular study of how numerically close the sentiment values are will require a better sentiment analysis tool. A custom lexicon that pertains to this data needs to be created to entirely understand what is going on. This lexicon could be used to train a large language model to understand radical language and give us more accurate sentiment values.

Further, for intra-thread, we can examine whether mixed sentiment con-

nections arise from agreement or disagreement, such as when someone posts a negative comment and the replier responds with positive sentiment or vice versa. This can be done with stance detection models that can tell if two pieces of text are in agreement or not. We are in the process of integrating them into our Graph-Based Approach to Studying the Spread of Radical Online Sentiment.

The data set we used was intentionally limited, but for this work and our models to have real-world usage, we will have to use much larger and more mixed data sets from sites like Twitter, Reddit, or 4Chan. Our limited data also makes our models very limited; though our models were only built to show the significance of our graphed-based feature set, there is much work to be done to improve them. We could combine the NLP features used in previous models with our novel graph-based features, we can train our models on much larger and more generalized data sets, and we could train our model on more partial graphs for it to have better predictive power or create an entirely separate model to make predictions.

Once our method and models are further refined, the end goal would be to use them to study the effects of sentiment spread on real-world outcomes, such as political mobilization or consumer behavior. By analyzing the impact of sentiment spread on these outcomes, we can better understand the social and economic implications of online sentiment.

6.3 Conclusion

This thesis explores a novel approach to studying the spread of online extremist sentiment. Using social network analysis and natural language processing, we examine the community dynamics of extremist comment threads and model the contagion of sentiment. We first use contact graphs to test the hypothesis that well-connected members share a similar sentiment. The results showed that around 73% of connected members shared similar sentiments within comment threads, and approximately 64% of connected comment threads shared a similar sentiment on the inter-thread level.

After performing a centrality study on all of our intra-thread graphs, we found that the original poster was the most central node and, therefore, the most influential member in 57% of our comment threads, using information centrality as our metric. Additionally, we found that the mean sentiment of a thread tends to match the sentiment of the original poster, with the mean difference between the original poster sentiment and the overall sentiment of a thread being 9.8×10^{-4}.

Temporal analysis or our threads backed up our previous findings. For the majority of our threads, the mean sentiment of the thread regressed to the sentiment of the original poster over time. The original poster also dominated as the central and most influential poster throughout time, being the most central poster 40% of the time on average, with the rest of the time being split among the rest of the members.

Once we gained an understanding of the behavior of our threads, we created a classification model that could determine the sentiment of a member in a thread based on the other members they interacted with. Using connected node sentiment, degree of connection, and centrality of both neighboring nodes and the target node, we were able to classify the sentiment of a target node with 87% accuracy. We then used our classifier as a contagion model that used an existing set of seed nodes to classify the sentiment of new nodes added to the graph. This contagion model could classify the sentiment of the next N nodes in the graph with 72%; starting with 5 known nodes.

This thesis proves the validity of using online community dynamics as a tool for the study of online hate. Traditional study of online hate only studying the language being used needs to incorporate social network features in order to obtain a full understanding of the online space. Incorporating the current research on understanding extremist language with the knowledge of community dynamics found in this thesis is the next step. Additionally, stance detection and a custom-trained large language model would increase the validity and utility of this work.

7 Appendix

List of Figures

List of Tables

References

[1] Varennes, UN Human Rights Council, "Recommendations made by the Forum on Minority Issues at its 13th session on the theme "Hate speech, social media and minorities"," 2021.

[2] G. Davies, "Radicalization and Violent Extremism in the Era of COVID-19," *The Journal of Intelligence, Conflict, and Warfare*, vol. 4, pp. 149–152, May 2021.

[3] M. Hamm and R. Spaaij, "Lone Wolf Terrorism in America: Using Knowledge of Radicalization Pathways to Forge Prevention Strategies, 1940-2013: Version 1," 2017. Type: dataset.

[4] A. Waqas, J. Salminen, S.-g. Jung, H. Almerekhi, and B. J. Jansen, "Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate," *PLOS ONE*, vol. 14, p. e0222194, Sept. 2019.

[5] O. Araque and C. A. Iglesias, "An Approach for Radicalization Detection Based on Emotion Signals and Semantic Similarity," *IEEE Access*, vol. 8, pp. 17877–17891, 2020.

[6] J. Torregrosa, G. Bello-Orgaz, E. Martínez-Cámara, J. D. Ser, and D. Camacho, "A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges," *Journal of Ambient Intelligence and Humanized Computing*, Jan. 2022.

[7] E. Ferrara, "Contagion Dynamics of Extremist Propaganda in Social Networks," *SSRN Electronic Journal*, 2017.

[8] A. Bermingham, M. Conway, L. McInerney, N. O'Hare , and A. F. Smeaton, "Combining Social Network Analysis and Sentiment Analysis to Explore the Potential for Online Radicalisation," (Athens, Greece), pp. 231–236, IEEE, 2009.

[9] G. Davies, "Radicalization and Violent Extremism in the Era of COVID-19," *The Journal of Intelligence, Conflict, and Warfare*, vol. 4, pp. 149–152, May 2021.

[10] J. Mothe, M. Z. Ullah, G. Okon, T. Schweer, A. Juršėnas, and J. Mandravickaitė, "Instruments and Tools to Identify Radical Textual Content," *Information*, vol. 13, p. 193, Apr. 2022.

[11] "Dark Web Forums: AZSecure-data.org."

[12] I. Pete, J. Hughes, Y. T. Chua, and M. Bada, "A Social Network Analysis and Comparison of Six Dark Web Forums," in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, (Genoa, Italy), pp. 484–493, IEEE, Sept. 2020.

[13] R. Kumar, J. Caverlee, and H. Tong, eds., *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: ASONAM 2016 : San Francisco, CA, USA, August 18-21, 2016*. Piscataway, New Jersey: IEEE, 2016. OCLC: 972638379.

[14] J. Zhang, J. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, D. Taraborelli, and N. Thain, "Conversations Gone Awry: Detecting Early Signs of Conversational Failure," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 1350–1361, Association for Computational Linguistics, 2018.

[15] J. Hessel and L. Lee, "Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features," in *Proceedings of the 2019 Conference of the North*, (Minneapolis, Minnesota), pp. 1648–1659, Association for Computational Linguistics, 2019.

[16] "What is the Deep and Dark Web?," Apr. 2023.

[17] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora," in *Proceedings of the*

2016 Conference on Empirical Methods in Natural Language Processing, (Austin, Texas), pp. 595–605, Association for Computational Linguistics, 2016.

[18] M. Ali and S. Zannettou, *Analyzing Antisemitism and Islamophobia using a Lexicon-based Approach*. US: ICWSM, June 2022.

[19] J. Torregrosa, J. Thorburn, R. Lara-Cabrera, D. Camacho, and H. M. Trujillo, "Linguistic analysis of pro-ISIS users on Twitter," *Behavioral Sciences of Terrorism and Political Aggression*, vol. 12, pp. 171–185, July 2020.

[20] N. Li and D. D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support Systems*, vol. 48, pp. 354–368, Jan. 2010.

[21] M. J. Keeling and K. T. Eames, "Networks and epidemic models," *Journal of The Royal Society Interface*, vol. 2, pp. 295–307, Sept. 2005.

[22] Rasim Alguliyev, Ramiz Aliguliyev, and Farhad Yusifov, "Graph modelling for tracking the COVID-19 pandemic spread," *Journal of Theoretical, Clinical and Experimental Morphology*, vol. 3, pp. 1–14, Feb. 2021.

[23] A. Majeed and I. Rauf, "Graph Theory: A Comprehensive Survey about Graph Theory Applications in Computer Science and Social Networks," *Inventions*, vol. 5, p. 10, Feb. 2020.

[24] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, (Lyon, France), pp. 913–922, ACM Press, 2018.

[25] A. Disney, "PageRank centrality & EigenCentrality," Jan. 2020.

[26] C. Amrit and J. ter Maat, "Understanding Information Centrality Metric: A Simulation Approach," Dec. 2018. arXiv:1812.01292 [cs, stat].

[27] M. Z. Al-Taie and S. Kadry, "Information Diffusion in Social Networks," in *Python for Graph and Network Analysis*, pp. 165–184, Cham: Springer International Publishing, 2017.

[28] Jie Tang and A. C. M. Fong, "Sentiment diffusion in large scale social networks," in *2013 IEEE International Conference on Consumer Electronics (ICCE)*, (Las Vegas, NV), pp. 244–245, IEEE, Jan. 2013.

[29] N. O. Hodas and K. Lerman, "The Simple Rules of Social Contagion," *Scientific Reports*, vol. 4, p. 4343, Mar. 2014.

[30] D. Küçük and F. Can, "Stance Detection: A Survey," *ACM Computing Surveys*, vol. 53, pp. 1–37, Jan. 2021.

[31] A. ALDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, p. 102597, July 2021.

[32] "Getting started with the built-in BERT algorithm | AI Platform Training."

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019.

[34] A. Sinha, S. Kedas, R. Kumar, and P. Malo, "Sentfin 1.0: Entity-aware sentiment analysis for financial news.," *Journal of the Association for Information Science and Technology*, vol. 73, pp. 1314–1335, Sept. 2022.

[35] P. J. Carrington, J. Scott, and S. Wasserman, eds., *Models and methods in social network analysis*. No. 27 in Structural analysis in the social sciences, Cambridge ; New York: Cambridge University Press, 2005.

[36] N. Sun, P. P.-L. Rau, and L. Ma, "Understanding lurkers in online communities: A literature review," *Computers in Human Behavior*, vol. 38, pp. 110–117, Sept. 2014.