BULLYNET:

UNMASKING CYBERBULLIES ON SOCIAL NETWORKS

by
Aparna Sankaran

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

Boise State University

December 2019

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Aparna Sankaran

Thesis Title: BullyNet: Unmasking Cyberbullies on Social Networks

Date of Final Oral Examination: 25 October 2019

The following individuals read and discussed the thesis submitted by student Aparna Sankaran, and they evaluated the presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Gaby Dagher, Ph.D.	Chair, Supervisory Committee
Bogdan Dit, Ph.D.	Member, Supervisory Committee
Min Long, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Gaby Dagher, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

In loving memory of my father

ACKNOWLEDGMENTS

I would first like to thank my thesis advisor Dr. Gaby Dagher. The door to Prof. Dagher's office was always open whenever I ran into trouble or had a question about my research or writing. He consistently allowed this paper to be my own work while steering me in the right direction whenever it was needed. His advising and technical expertise were very helpful throughout my graduate study and made the research process comfortable and smooth. I would also wish to express my gratitude to the committee members, Dr. Bogdan Dit and Dr. Min Long for discussions and suggestions on my thesis. I am fortunate to be a part of Boise State University with an incredible group of fellow students. I am grateful to them for their support throughout my graduate studies, especially to Hannah Johnson.

I must express my profound gratitude to my husband, Srinath, for providing me with unfailing support and continuous encouragement throughout my graduate study and through the process of researching and writing this thesis. This accomplishment would not have been possible without him. I would like to express deepest gratitude to my always supportive mother-in-law and father-in-law who were there during the difficult times to help me through the graduate studies. Special mention goes to our two wonderful kids - Brihad Hari and Athri Sankar for bringing joy into our lives and for being very co-operative during my graduate study. I thank my sister, grandmother and grandfather for their love and affection. Last but not the least, I thank my mom for her moral and emotional support throughout my graduate studies and for instilling the importance of academics right from my childhood, being a teacher herself.

ABSTRACT

Social media has changed the way people communicate with each other, and consecutively affected people's ability to empathize in both positive and negative ways. One of the most harmful consequences of social media is the rise of cyberbullying, which tends to be more sinister than traditional bullying given that online records typically live on the internet for quite a long time and are hard to control. In this thesis, we present a three-phase algorithm, called BullyNet, for detecting cyberbullies on Twitter social network. We exploit bullying tendencies by proposing a robust method for constructing a cyberbullying signed network. BullyNet analyzes each tweet to determine its relation to cyberbullying, while considering the context in which the tweet exists in order to optimize its bullying score. We also propose a centrality measure to detect cyberbullies from a cybebullying signed network, and we show that it outperforms other existing measures. We evaluate our method on a dataset of 5.6 million tweets we synthesized and labeled. Our experimental results show that the proposed BullyNet algorithm can detect cyberbullies with high accuracy, while being scalable with respect to the number of tweets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SSN – Signed Social Network

SA – Sentiment Analysis

CS – Cosine Similarity

BAD – Bias and Deserve

A&M – Attitude and Merit

LIST OF SYMBOLS

$\mathcal{UID}$ denotes User ID

$\mathcal{RID}$ denotes Reply ID

$\mathcal{MID}$ denotes Mentions ID

$\mathcal{SID}$ denotes Source ID

$\mathcal{DID}$ denotes Destination ID

e_{ij} denotes directed edge from i to j

$\mathcal{B}$ denotes Bullying Signed Network

β denotes coefficient of sentiment analysis

γ denotes coefficient of cosine similarity

c_i denotes i^{th} conversation

g_{c_i} denotes graphical representation of conversation i

I_{uv} denotes bullying indicator value from node u to v

G_c denotes total conversation graphs

S_{uv} denotes bullying score of a node u to v

w_{uv} denotes edge weight from noe u to v

α denotes coefficient for the response tweet

$in(v)$ denote the set of all incoming edges to node v

$out(u)$ denote the set of all outgoing edges from node u

$M^n(u)$ denotes merit of a node u for n^{th} iteration

$A^n(u)$ denotes attitude of a node u for n^{th} iteration

u denotes bullying user

s denotes confidence score of the bullying user

CHAPTER 1

INTRODUCTION

The Internet has created never before seen opportunities for human interaction and socialization. In the past decade, social media, in particular, has had a popularity explosion. From MySpace to Facebook, Twitter, Flickr, and Instagram, people are connecting and interacting in a way that was previously impossible. The widespread usage of social media across people from all ages created a vast amount of data for several research topics, including recommender systems [54], link predictions [37], visualization, and analysis of social networks [6].

1.1 Motivation

While the growth of social media has created a good platform for communications and information sharing, it has also created a new platform for malicious activities such as spamming [25], trolling [8], and cyberbullying [32]. According to the Cyberbullying Research Center (CRC) [44], cyberbullying occurs *when someone uses the technology to send messages to harass, mistreat or threaten a person or a group*. Unlike traditional bullying where aggression is a short and temporary face to face occurrence, with cyber-bullying, the bullying messages are always present online, can be accessed worldwide, and are often irrevocable. A series of surveys done by the CRC shows that the rates of cyberbullying among youth have increased over the past nine years, going from 18.8% in

2007 to 33.8% in 2016, where most of the victims suffered from either *hurtful comments* or *rumor spreading*.

Laws about cyberbullying and how it is handled differ from one place to another. For example, in the United States, the majority of the states incorporate cyberbullying into their bullying laws, and cyberbullying is considered a criminal offence in most of them [9]. On the other hand, in Europe, all 47 member states of the Council of Europe have adopted the *Charter on Education for Democratic Citizenship and Human Rights Education* [13], which mandates them to fight all forms of discrimination and violence including cyberbullying.

In order to identify cyberbullies in social media, we first need to understand how social media can be modeled. The most common way of modeling social media networks is to represent it as a graph, where nodes correspond to users and edges correspond to communications and/or relations between the users. When each edge in the graph is directed and assigned a value (weight) from the range [-1, 1], then the graph is called a *signed network*, as illustrated in the example in Figure 1.1.

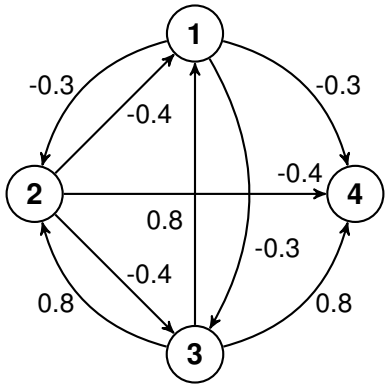


Figure 1.1: An example of a signed network.

The following is a formal definition of a signed social network.

Definition 1.1.1. A signed social network (SSN) is a directed, weighted graph $G = (V, E, W)$, where V is the set of users and $E \subseteq V \times V$ is the set of edges with an edge weight $w \in W$ in the range of [-1,1].

1.2 Challenges & Concerns

Mining social media networks to determine cyberbullies imposes several challenges and concerns. It is typically hard to accurately interpret user's intentions and meanings in social media based merely on their messages (e.g. posts, tweets, comments), which are typically short, use slang languages, and may include multimedia contents such as pictures and videos. For example, Twitter limits its users' messages to 140 characters, which could be a mix of text, slangs, emojis, and gifs. As a result, it is hard to correctly determine the sentiment of a message. In addition, bullying could be hard to detect if the bully chooses to disguise it through techniques such as sarcasm or passive-aggression. Furthermore, the large size and dynamic and complex structure of social media networks makes it difficult to identify cyberbullies. For example, in Twitter, hundreds of millions of tweets are sent every day on the social network platform. There are several works in the literature concerning detecting malicious users from unsigned networks with positive edge weights, including community detection [57], node classification [2] and link prediction [37]. On the other hand, methods that analyze signed social networks are scarce [55].

In this thesis, we study the problem of cyberbullying in social media and propose an approach for efficiently detecting cyberbullies in the Twitter social network. Our solution consists of three parts. Our intuition is that each tweet should be evaluated not only based on its contents, but also based on the context in which it exists. We call such a context a

conversation, which is a set of tweets between two or more people exchanging information about a certain subject. First, for each conversation, a conversation graph is generated based on the sentiment and bullying words in the tweets. Second, we compute the bullying score for each pair of users in a conversation graph, and then combine all graphs to create a signed social network called bullying signed network ($\mathcal{B}$). The inclusion of negative links can bring out information that would otherwise be missed with only positive links [33]. Finally, we propose a centrality measure called attitude & merit ($A\&M$) to detect bullying users from the signed network $\mathcal{B}$.

1.3 Thesis Statement

The objective of this thesis is to answer the following question: **How to identify cyberbullies effectively on Twitter?**

Given the Twitter network, let $T = \{t_1, t_2, ...t_{|T|}\}$ be a dataset of tweet objects, where each tweet object t includes the text of a tweet, source ID, destination ID, date of creation, user name, reply name, and mentions. The objective of this research is to propose an approach that utilizes a cyberbullying centrality measure in order to accurately identify the set of users $\{u_1, u_2, \dots\}$ in T who exhibit bullying behaviour such that:

1. The proposed approach assigns a bullying confidence value s_i to each identified user u_i.

2. The proposed approach is efficient and scalable.

1.4 Thesis Contribution

Our main contributions are organized as follows:

- Collected, preprocessed and labelled the Twitter dataset.
- Proposed an efficient algorithm for detecting cyberbullies in Twitter.
  - Built conversation.
  - Constructed Bullying Signed Network.
  - Proposed Attitude and Merit Centrality.
- Experimented on 5.6 million tweets collected over 6 months. The results show that our approach can detect cyberbullies with high accuracy, while being scalable with respect to the number of tweets.

1.5 Organization of the Thesis

The thesis is organized as follows:

- Chapter 2 introduces the preliminaries required for the protocols such as sentiment analysis to analyze the sentiment of the message, cosine similarity to measure similarities between the messages and centrality measures for analyzing the signed network .
- Chapter 3 explains the in-depth literature review of the previous cyberbullying detection categorized into detecting messages and detecting users, signed network used to understand and represent the social media and measures that are used to analyze nodes and edges in signed networks.
- Chapter 4 is about the how the problem is formulated.
- Chapter 5 provides details of the proposed bully finding three phase algorithm for detecting cyberbullies in the Twitter social network.

- Chapter 5 provides the algorithm analysis of our proposed approach that includes the proof of convergence of the centrality measure and complexity analysis.
- Chapter 6 illustrates the experiments and performance evaluation of our proposed solution.
- Chapter 7 concludes our work and discusses the future work.

CHAPTER 2

BACKGROUND

This chapter introduces and defines the building blocks and preliminaries that help establish the foundational knowledge required to better understand the proposed work. They are Sentiment Analysis, Cosine Similarity, and Centrality Measures.

2.1 Sentiment Analysis

Sentiment analysis (SA) is the process of analyzing the sentiment of a message based on the user's opinion, attitude, and emotion towards an individual. Depending on the analysis, the polarity of the text is classified into positive, negative or neutral. The sentiment reflects feeling or emotion while emotion reflects attitude. It was argued by Plutchik [46] that there are eight basic and prototypical emotions which are joy, sadness, anger, fear, trust, disgust, surprise, and anticipation.

SA is a classification technique which derives opinion and attitude from the message and formulates a sentiment score reflecting the sentiment-based insights of the text. It has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task [58, 43], it has been handled at the sentence level [24, 30] and then at the phrase level [61, 1]. A very broad overview of the existing work was presented by Medhat *et al.* In their survey [39], the authors describe existing techniques, approaches and their application. The techniques are classified in two ways.

The first approach is the machine learning technique (ML), which relies on specific ML algorithms to solve the SA as a regular text classification problem making use of syntactic and/or linguistic features. It uses various classifiers such as Linear, Decision Tree, Naive Bayes etc, to identify the expressed sentiment. The second approach is Lexicon-based, which uses a variety of words annotated by polarity score, to decide the general assessment score of a given text. The Lexicon-based approach is further divided into the dictionary-based approach and corpus-based approach, which uses statistical or semantic methods to find sentiment polarity.

Depending on the platform and programming languages, there are different libraries or tools available to determine the sentiment of the message content, which includes sarcasm, emoji, images etc. Some of the them are: VADER, TextBlob, Python NLTK etc. VADER (Valence Aware Dictionary and sEntiment Reasoner) [26] is used in our work, which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is a combination of sentiment lexicon with consideration for five general rules that embody grammatical and syntactical conventions for expressing and emphasizing sentiment intensity. It contains a list of lexical features that are generally labeled according to their semantic orientation as either positive or negative. It performs well with emojis, emoticons, slangs and acronyms in a sentence.

Example 2.1.1. Let's consider the following sentence:

I just got a call from my boss - does he realise it's Saturday? smh :(

where the sentence polarity score is presented in Table 2.1. The Positive, Negative and Neutral scores represent the proportion of text that falls in these categories. This table shows that the sentence was rated as 32% Positive, 68% Neutral and 0% Negative. All these should add up to 100%. The Compound score is a metric that calculates the sum of all the lexicon ratings, which have been normalized between -1 (most extreme negative)

and +1 (most extreme positive). Since the Compound score for this message is -0.63, this mean, the message has a very high negative sentiment. □

Table 2.1: VADER Polarity score

Sentiment Metric	Score
Negative	0.321
Neutral	0.679
Positive	0.0
Compound	-0.6369

2.2 Cosine Similarity

Cosine Similarity (CS) [21] is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. When text are represented as term vectors, the similarity of two texts corresponds to the correlation between the vectors. As a result, the cosine similarity is non-negative and bounded between [0,1]. It is one of the most popular similarity measures applied to texts or documents, such as in numerous information retrieval applications and clustering.

Let X and Y be two vectors for comparison. Using the cosine measure as a similarity function, we have

$$CS(\vec{X}, \vec{Y}) = \frac{\vec{X}.\vec{Y}}{\parallel \vec{X} \parallel . \parallel \vec{Y} \parallel}$$

The measure computes the cosine of the angle between vectors X and Y. A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match. The closer the cosine value is to 1, the smaller the angle and the greater the match between vectors.

In this proposal we focus on Twitter, since some tweets contain curses or insulting words, these words are good indications of the existence of bullying. Therefore, we select a list of insulting words which are commonly used in Twitter and also in some external linguistic resources, which are taken as insulting seeds. This list contains words indicating curse or negative emotions such as *nigga, bitch, slut etc.* and are compared with individual tweets with cosine similarity to compute a score. In this context, each tweet and insulting seeds are represented as vectors, where each vector has the word frequencies.

2.3 Centrality Measures

Centrality is a measure in a network that is used to identify the most important vertices and also to determine how one vertex affects others in a network. Given a signed network (SN) with a graph $G = (V, E, W)$, the centrality measure of a node is a function $F : V \rightarrow \mathbb{R}$ that assigns a numerical value to each vertex of a network according to its influence on the others. The importance of a node is determined by how high the score is within a network and also defined by the type of the network. It could be identified as an effective person in a social network or key infrastructure nodes in the urban networks. Some of the centrality measures used in social networks are: PageRank, Modified Pagerank, Hyperlink-Induced Topic Searc (HITS), Modified HITS, PageTrust and Bias and Deserve (BAD).

PageRank [7] is a link analysis algorithm and it was designed for unsigned networks. It measures the transitive influence or connectivity of nodes, i.e., the PageRank assigns a score to a node based on its connections, and its connections' connections. It was originally developed to rank websites in Google's search results. The PageRank of a node u is defined

as:

$$PR(u) = \frac{1-\alpha}{N} + \alpha \sum_{v \in out(u)} \frac{PR(v)}{|out(v)|}$$

Here, α is a "damping factor" (usually 0.85) which captures the probability that a user arrives at a web page by following links. $out(v)$ is the set of all outgoing edges of the node v. For signed network, *Modified Pagerank* has been proposed in [49] to take into account both positive and negative links. It is computed as the difference between PR^+ and PR^-.

Hyperlink-Induced Topic Search (HITS) [31] was originally proposed to analyze the link structure in the World Wide Web (WWW). It has two metrics known as Authority and Hub, its values are defined in terms of one another in a mutual recursion. An Authority value is computed as the sum of the scaled hub values that point to that page. A Hub value is the sum of the scaled authority values of the pages it points to. The HITS of a node u is defined as:

$$A(u) = \sum_{v \in out(u)} H(v)$$

$$H(u) = \sum_{v \in out(u)} A(v)$$

In *Modified HITS* [49], a signed network similar to Modified PageRank, the overall authority value is computed as $A(u)^+$ - $A(u)^-$, where $A(u)$ denotes the corresponding authority values for the node u.

PageTrust [29] extends the PageRank algorithm to include the negative links in a network. It multiplies the PageRank equation with a heuristic correction factor in an effort to account for negative links. The HITS of a node u is defined as:

$$\sum_{v \in S} P_{uv}^{t+1} . \sum_{v \in N/S} (1 - P_{uv}^{t+1})$$

where S is the list of nodes $i_1, i_2, ..$ and N is number of distrust nodes.

Bias and Deserve (BAD) is a centrality measure proposed by Mishra and Bhattacharya [41] which is closely related to our paper, designed for signed networks to compute the bias or the truthfulness of a user in trust based networks and shows that there are users who have a propensity to trust or distrust other users. The 'Bias' of a node reflects the expected weight of an outgoing edge, while 'Deserve' reflects the expected weight of an incoming edge from an unbiased node. The corresponding equations are:

$$Bias^{t+1}(u) = \frac{1}{2|out(u)|} \sum_{v \in out(u)} \{W_{uv} - Deserve^{t+1}(v)\}$$

$$Deserve^{t+1}(u) = \frac{1}{|in(u)|} \sum_{v \in in(u)} \{W_{vu}(1 - X^t(vu))\}$$

where $X^t(uv) = max(0, Bias^t(v) \times W_{vu})$, $out(u)$ is the set of all outgoing edges from the node u and $in(v)$ is the set of all incoming edges to the node v.

CHAPTER 3

LITERATURE REVIEW

In this section, we review the literature examining several areas related to our work which is cyberbullying detection, signed networks, and bully detection using signed networks.

3.1 Cyberbullying Detection

Much work has been done over the past decade in the area of cyberbullying detection. There have been two broad approaches in identifying bullies - one aims to detect bullying messages [65, 50, 23, 63, 14, 15, 16], while the other approach is to detect the cyberbullies responsible for the messages [51, 18, 10, 11].

The first approach is to determine bullying messages, some used text-based analytics, and others used a mix of text and user features. Zhao *et al.* [65] proposed a text based Embeddings-Enhanced Bag-of-Words (EBoW) model that utilizes a concatenation of bullying features, bag-of-words, and latent semantic features to obtain a final representation, which is then passed through a classifier to identify cyberbullies. Xu *et al.* [63] used textual information to identify emotions in bullying traces, as opposed to determining whether or not a message was bullying. Singh *et al.* [50] proposed a probabilistic socio-textual information fusion for cyberbullying detection. This fusion uses social network features derived from a 1.5 ego network and textual features, such as density of bad words and

part-of-speech-tags. Hosseinmardi *et al.*, [23] used images and text to detect cyberbullying incidents. The text and image features were gathered from media sessions containing images and the corresponding comments, which were then fed into various classifiers. Dadvar *et al.*, [14] proposed an approach to improve the performance of detection tools for cyberbullying incidents with supervised classification on a combination of content-based, cyberbullying-specific and user-based features. While Dinakar *et al.*, [16] applied a set of features similar to [14] along with some other features which were specific to the topic of the videos. The authors also address detection [15] of potentially bullying messages and the way to intervene by notifying participants and network moderators, managing message access, and offering targeted educational material.

The second approach was aimed at identifying the person behind the cyberbullying incidents. Squicciarini *et al.* [51] used MySpace data to create a graph, which integrated user, textual, and network features. This graph was used to detect cyberbullies and predict spreading of bullying behavior through node classification. Chen *et al.* [11] proposed the use of a lexical syntactic feature approach to detect offensive content and potential offensive users. They observe the contribution of pejoratives/profanities and obscenities in determining offensive content by considering the writing style of the users. Galn-Garca *et al.* [18] used supervised machine learning to detect the real users behind troll profiles on Twitter, and demonstrated the technique in a real case of cyberbullying. Similar to the objectives of the present work, Chatzakou *et al.* [10] investigated user features that can be utilized to enhance the detection and classification of bullying by employing supervised machine learning to classify Twitter users into four classes: bully, aggressive, spam, and normal behaviour, using network based features like user id, text etc.

However, the above papers focus mainly on how offensive the content of the message is rather than specifying how the message is viewed by the receiver. They also do not consider

the previous response that was sent to the sender to identify whether the user is a bully or not.

3.2 Signed Networks

This section reviews the previous work done on signed networks. Some of the papers describing signed networks in social media are [35, 53, 55, 62, 32]. Leskovec *et al.* [35] reviewed the Balance and Status theory and its relation to social media and proposed a modified status theory that more accurately reflects patterns found in signed networks in social media. Signed networks themselves are not a new idea, but the application and analysis of them are new. With the widespread presence of social media networks, signed network analysis has evolved from developing and measuring theories to mining tasks. Tang *et al.* have done a broad survey of mining signed networks in social media [55]. The authors classify tasks of mining signed networks into applications, links and nodes.

Some of the application-oriented tasks in signed networks are, data classification [66] where Zhu *et al.* propose a simple approach to combine the information of content and links for web page analysis, mainly using classification application. Long *et al.* proposed a model on data clustering [38] in Multi-Type Relational Data with various structures of different types of data objects, dimensionality reduction and noise removal leading to better embeddings. Information diffusion in applications such as effective viral marketing [28], and social recommender systems [56].

The link-oriented tasks are links among nodes, which aim to reveal detailed understandings of links. The papers [17, 5, 64, 34, 12] list some of the supervised classification methods that use the existence of links as labels in link prediction, while the unsupervised methods are usually based on certain topological properties of signed networks,

such as similarity and propagation based methods [52, 20, 67]. In link prediction [34], triangle-based features are extracted based on balance theory to improve link prediction. Since signed social networks are usually very sparse and most users have no triangle-based features, Chiang *et al.* [12] developed an algorithm based on any quantitative social imbalance measure of a signed network. Other types of features are also used for the problem of link prediction like user interaction features [17] and review-based features [5]. [52] provided friend recommendations, by performing multi-way spectral clustering based on the Laplacian matrix for signed networks. Guha *et al.* [20] developed trust propagation schemes to trust or distrust each individual and predict trust between any two people in the system.

The node-oriented tasks are further classified into community detection, node embedding and node classification. In Community detection, the group of users are identified, where users are densely connected by positive and negative links in the groups [3, 36]. Bogdanov *et al.* [3] proposed an end-to-end framework for analysis of community relations based on collaboration activities and Li *et al.* [36] developed an Modularity-Based Algorithm to detect communities in signed networks. The node embedding is to learn low-dimensional vector representations for nodes of a given social network. Wang *et al.* observed attribute similarity of users with positive, negative and no links and proposed an framework [60] by incorporating the extended structure balance theory and the relationship between user links attributes. This proposal focuses on node classification with signed social networks, where a user information such as demographic values, interest beliefs, or other characteristics plays an important role in social media. According to [2] node classification algorithms in unsigned social networks can be mainly divided into local classifier methods and random walk based methods. Although there are many algorithms focused on node classification in unsigned social networks, the existing literature on signed social

networks is less. Getoor and Diehl performed an extensive study on node classification [19] i.e., on nodes' attributes, their links to other nodes and the attributes of these linked nodes. Tang *et al.* observed both positive and negative links and proposed an approach [53] to mathematically model both independent and dependent information from the links on node classification.

3.3 Measures to analyze Singed Network

Over the last few years, a number of algorithms have been designed to prioritize the set of nodes on unsigned networks that cannot deal with negative values directly by ignoring negative links or zero the entries corresponding to the negative links [22, 27, 47]. In recent times, few measures have been designed for signed network analysis with both positive and negative links [49, 29, 4, 62, 59]. Most of these methods are based on modifications of the PageRank [7] or HITS [31] centrality which accounts for negative weights on the links.

Exponential ranking in [59] was designed for ranking nodes in signed networks by heuristically using an exponential variation of the PageRank equations to deal with negative links. A Modified PageRank has been proposed in [49] to take into account both positive and negative links. In particular, they apply PageRank separately for both links and obtain the difference between them. Similarly, the modified HITS in [49] iteratively computes the hub (positive links) and authority (negative) scores separately. However, some of these measures consider every node in isolation when computing the centrality and also completely ignore the interactions between the positive and negative links. For example, if a network contains a large number of positive links relative to the negative links, and vice versa, it should affect the opinion values. Mishra and Bhattacharya [41] proposed bias and prestige measures based on HITS algorithm in a trust network where, the prestige of a node

depends on the opinions of other nodes whereas trustworthiness of a node depends on how a node gives correct opinion about other nodes. Though there is interaction between the links, it is not effective for identifying bullies in the network.

The papers [62, 32, 42], are aimed at detecting trolls in a signed network. Ortega *et al.* [42] aims to detect trolls in a social network by computing a rank for the users based on the trustworthiness. The paper [62] proposed a method for ranking nodes to identify trolls that models the probability of trustworthiness of individual data without using a modified version of the PageRank algorithm. Kumar *et. al* [32] proposed an iterative algorithm involving new decluttering operations and various centrality measures to detect trolls. Unlike the proposed method in this paper, the authors begin their process with an already created signed network.

Table 3.1 summarizes the features of the representative approaches, including our proposed protocol.

Table 3.1: Comparative evaluation of main features in related approaches including our proposed approach

Approach	**Malicious Activity**			**Attributes**				**Signed Network**		**Dataset**			
	Cyberbullying		**Other**	**Content-based**	**Context-based**	**User-based**	**Network-based**	**Yes**	**No**	**Twitter**	**Youtube**	**Slashdot**	**Instagram**
	Message	**User**											
Zhao *et al.* [65]	●			●					●	●			
Xu *et al.* [63]	●			●					●	●			
Hosseinmardi *et al.*, [23]	●			●					●				●
Dadvar *et al.*, [14]	●			●					●		●		
Dinakar *et al.*, [16]	●			●		●			●		●		
Squicciarini *et al.* [51]		●		●	●	●	●		●	●			
Chen *et al.* [11]		●		●					●		●		
Galn-Garca *et al.* [18]		●		●		●			●	●			
Chatzakou *et al.* [10]		●		●		●	●		●	●			
Mishra & Bhattacharya [41]			●				●	●				●	
Kumar *et. al* [32]			●				●	●				●	
Wu *et. al* [62]			●				●	●				●	
Ortega *et al.* [42]			●				●	●				●	
Our proposed protocol		●		●	●		●	●		●			

CHAPTER 4

PROBLEM FORMULATION

In this chapter, a Twitter social network is represented as a directed, weighted graph $G = (U, E)$ with U being the set of users (represented as nodes) and E being the set of tweets T between users (represented as edges). Each user $u \in U$ has a set of features including an ID, the number of followers, the number of friends, and the number of the tweets that they sent. For any tweet, $t \in T$, there also exist certain features, which include a source ID ($\mathcal{SID}$), destination ID ($\mathcal{DID}$), the date of creation, a user ID ($\mathcal{UID}$), a reply ID ($\mathcal{RID}$) and mentions ($\mathcal{MID}$) if the tweet includes one. If a given @username is included in a tweet anywhere else but at the very start, Twitter interprets this as a mention and the user gets a notification that someone has mentioned them.

The notation e_{ij} represents a directed edge from node i to node j. The existence of an edge e_{ij} denotes an interaction from node i to node j. The edge is directional, but is not guaranteed to be reciprocated. For example, the existence of an edge e_{ij} does not guarantee the existence of an edge e_{ji}. The attributes of an edge, e_{ij} are a start node, an end node, and a weight.

From the above Twitter data, we extract conversations and build a directed weighted graph for each conversation $C = \{C_1, C_2, ...C_{|C|}\}$.

A conversation is a set of tweets between two or more users. More formally:

Definition 4.0.1. A conversation c is a set of time-ordered tweets $c = \{t_1, t_2, ...t_{|c|}\}$ such that:

1. The first tweet t_1 is the *initiator* tweet that starts the conversation, and can be of one of the two following types:
   - $\mathcal{DID}(t_1)$ =NULL, and either $\mathcal{MID}(t_1)$ or $\mathcal{RID}(t_1)$ is not null.
   - $\mathcal{DID}(t_1) \neq$ NULL, and $\forall t \subseteq T : \mathcal{SID}(t) \neq \mathcal{DID}(t_1)$.
2. All tweets in c satisfy the following:
   $\mathcal{SID}(t_i) = \mathcal{DID}(t_{i+1}) : 1 \leq i \leq |c| - 1.$ □

The goal is to output a list:
$L = \{(u_1, s_1), (u_2, s_2), ..., (u_{|L|}, s_{|L|})\}$ where u_i is a user (node) and s_i is a confidence value for the likelihood of user u_i being a bully.

CHAPTER 5

PROPOSED METHOD

This chapter first presents an overview of the proposed bully finding three phase algorithm for detecting cyberbullies in the Twitter social network, then elaborate the key step in each phase.

The objective of our solution is to identify bullies from raw Twitter data based on the context as well as the contents in which the tweet exists. Given a set of tweets T containing the Twitter features such as user ID, reply ID etc, our approach consists of three algorithms - (i) Conversation Graph Generation Algorithm, (ii) Bullying Signed Network Generation Algorithm and (iii) Bully Finding Algorithm. The first algorithm constructs a directed weighted conversation graph G_c by efficiently reconstructing the conversations from raw Twitter data while enabling a more accurate model of human interactions. The second algorithm constructs a bullying signed network $\mathcal{B}$ to analyze the behaviour of users in a social media. Finally, the third algorithm consists of our proposed attitude and merit centrality measures to identify bullies from $\mathcal{B}$. Figure 5.1 shows the process flow of BullyNet where the raw data is extracted from Twitter using Twitter API from which the conversation graph is constructed for each conversation using algorithm 5.1. Then from the conversation graphs, a bullying signed network is generated using algorithm 5.2. Finally the bullies from Twitter are identified by applying algorithm 5.3.

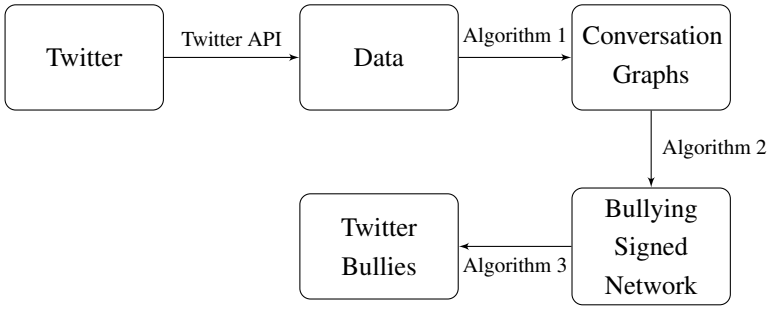


Figure 5.1: Protocol Flowchart

5.1 Algorithm 1 - Conversation Graph Generation

The conversation graph generation algorithm 5.1, is constructed from a set of tweets $T = \{t_1, \ldots, t_n\}$ to generate a directed weighted conversation graphs $G_c = \{g_{c_1}, \ldots, g_{c_m}\}$ for each conversations c_i, which is extracted from the tweets T. The graphs are represented as $G_c = (V, E)$ where V is the set of users involved in the conversation, E is the set of edges representing the tweets in the conversation, and each edge is assigned a bullying indicator value I as the edge weight which is in the range of $[-1, +1]$. When $I_{ij} = -1$, it indicates the negative interaction by i towards j and when $I_{ij} = 1$, it indicates the positive interaction by i towards j. The bullying indicator I, for each tweet is calculated based on sentiment analysis and cosine similarities. In Step 1, the tweets set T are sorted based on the creation time to reduce the time complexity, while searching for tweets based on $\mathcal{DID}$. Moreover, the set is sorted in a reverse-chronological order so that every $\mathcal{DID}$ of a tweet matches with only one $\mathcal{SID}$ of the remaining tweets. In Step 2, for each tweet t_i in T, the conversations are built by doing a binary search $\mathcal{DID}(t_i)$ with the $\mathcal{SID}$ of the remaining tweets. If a match is found as t' then, it is appended with t_i to form a new conversation. If binary search match is found with the already existing tweet in the conversation c_i then, t_i is appended to tweets in c_i.

Conversation Graph Generation Algorithm
Input: Set of tweets, $T = \{t_1, \ldots, t_n\}$
Output: Conversation graphs $G_c = \{g_{c_1}, \ldots, g_{c_m}\}$

1. Sort all tweets in T in reverse-chronological order based on date of creation.
2. For each tweet t_i in T, where $1 \leq i \leq |T|$:
   (a) If t_i does not belong to a conversation, then create a new conversation $c \in C$ and associate t_i with c.
   (b) If there is a tweet $t' \in \{t_i, t_{i+1}, \ldots, t_{|T|}\}$ where $\mathcal{DID}(t_i) = \mathcal{SID}(t')$ then associate t' with all t_i's conversations.
3. For each conversation $c_i \in C$:
   (a) Construct a conversation graph $g_{c_i} \in G_c$, where users are represented as nodes and tweets as edges.
   (b) For each edge $e = (u, v)$ in g_{c_i}:
      i. Compute the sentiment of the tweet (SA).
      ii. Compute the cosine similarity of the tweet with bullying bag of words (CS).
      iii. Calculate the bullying indicator I_{t_i} (weight) of the edge as follows: $I_{uv} = \beta * SA + \gamma * CS$
4. Return G_c

Algorithm 5.1: Conversation Graph Generation

Example 5.1.1. Figure 5.2 illustrates the conversation extracted from the set of tweets $T = \{t_1, \ldots, t_7\}$. First, the tweets are sorted in descending order i.e., $t_7, t_6, \ldots, t_1$. Next, $\mathcal{DID}(t_7)$ is searched with the $\mathcal{SID}$ of the remaining tweets (t_6 through t_1). A match is found in t_3 and conversation c_1 is formed. This process is repeated for each tweet. The conversations c_2 and c_3 are created with tweets $\{t_6, t_4\}$ and $\{t_5, t_2\}$ respectively. Since $\mathcal{DID}(t_4)$ and $\mathcal{DID}(t_3)$ match with the $\mathcal{SID}(t_1)$, the tweet t_1 is appended with t_4 and t_3.

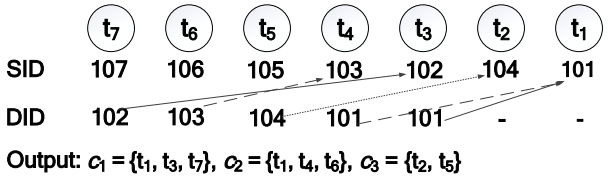


Figure 5.2: Matching tweets based on $\mathcal{DID}$ and $\mathcal{SID}$ to construct conversations. Given tweets $\{t_1, \ldots, t_7\}$, the output is three conversations: c_1, c_2, and c_3.

So, the final conversations are $c_1 = \{t_7, t_3, t_1\}$, $c_2 = \{t_6, t_4, t_1\}$ and $c_3 = \{t_5, t_2\}$ □

In Step 3, a directed, weighted graph $g_{c_i} = (V, E)$ is constructed for every conversation c_i where nodes V, represented as the users, and the edges E, represented as the tweets, directed from one user to another in a conversation. For every edge e, an edge weight is calculated as $I = \beta * SA + \gamma * CS$. This known as bullying Indicator which is in range of $[-1, +1]$. The sentiment analysis (SA) and cosine similarity (CS) is computed on the tweet (edge) to evaluate the emotion and behaviour of the user. The β and γ are constant, which will be determined by the experiment. In Step 4, the algorithm outputs the conversation graphs G_c.

Example 5.1.2. Figure 5.3 shows the Twitter example. From the algorithm 5.1, the conversation graphs are constructed as shown in Figure 5.4. It contains two conversation graphs shown with dashed blue edges and with solid red edges. The rounded number on the edges indicate the tweet order of that particular conversation. Figure 5.5a and 5.5b represent the two conversation graphs as g_{c_1} and g_{c_2} with the bullying indicator as the edge weight. With β and γ values as 0.9 and 0.1 which was determined by the experiment, the edge weight I_{31} i.e., the edge from $P3$ to $P1$, is calculated as -0.23. Similarly, the score of the other edges are calculated as shown in Figure 5.5a and 5.5b. □

P1 @P1 .28 Nov
President #Trump left the Christmas tree lighting quickly and unexpectedly. He is back at the White House, a full lid has been called. We don't know what's going on at this moment.

P2 @P2 .28 Nov
Replying to @P1
What is full lid?

P3 @P3 .28 Nov
No more news from WH today. Reporters can go home. What a moron to leave early?

P2 @P2 .28 Nov
Oh, thanks. I wonder why he left so suddenly!

P4 @P4 .28 Nov
Replying to @P1 @P5 @P3
Anyone noticed he hasn't looked well as of late?

P5 @P5 .28 Nov
Yup. He looks terrible

P4 @P4 .28 Nov
He looks like a patient at shadypines

P3 @P3 .28 Nov
He looks like a walking corpse

Figure 5.3: Sample conversation tweets

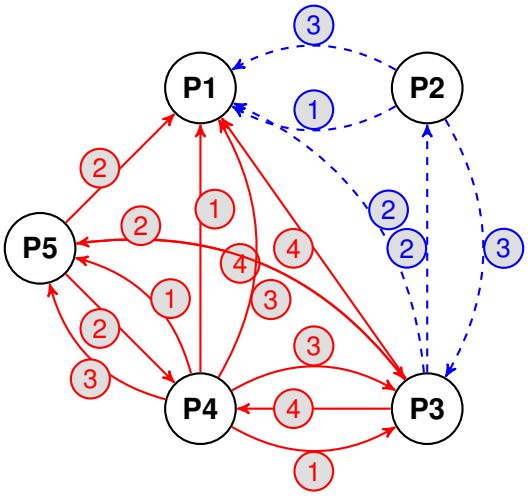
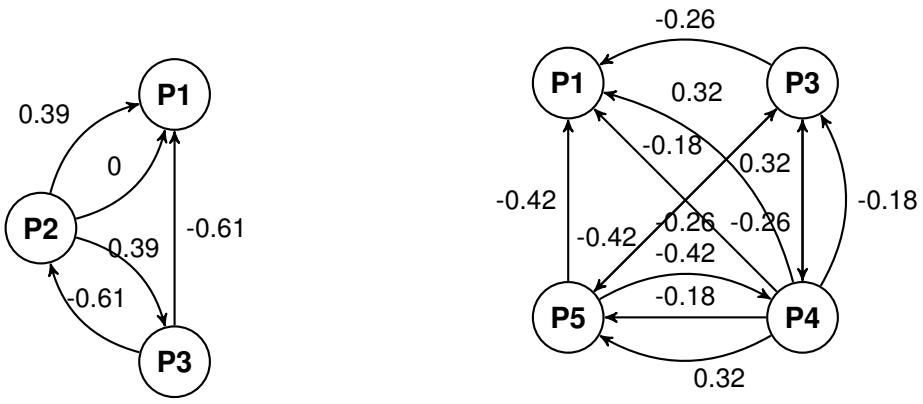


Figure 5.4: Conversations graph, blue - conversation 1, red - conversation 2



(a) g_{c_1} (b) g_{c_2}

Figure 5.5: Conversation Graphs with bullying indicators as edge weights.

5.2 Algorithm 2 - Bullying Signed Network Generation

In many real-world social systems, the relation between two nodes can be represented as signed networks with positive and negative links. Since this research focuses on identifying the bullying nodes in the network, the algorithm 5.2 is designed to determine the final outgoing edge weight, w_{ij} for the users in the conversation graphs G_c. In Step 1a, for every conversation graph g_{c_i}, a bullying score S is calculated for the users(nodes) in that graph based on the tweet order (sorted in ascending order). For an edge $e = (u, v)$, the bullying score S_{uv} is set to I_{uv} if the edge towards v is not a reply from u or else, the bullying score S_{uv} is calculated as $I_{uv} + ((I_{uv} - S_{vu}) * \alpha)$ where α is a constant which will be determined by the experiment. If there are more than one edge for a user with the same order then, after the bullying score is evaluated, an average bullying score is computed for the same set of order.

Example 5.2.1. Table 5.1 shows the bullying score calculation for the conversation graph g_{c_1} in Figure 5.5a. In order 1, the bullying score $S_{21} = I_{21} = 0$ since, the edge from $P2$ to $P1$ is not a reply edge. The user in the parenthesis represents to whom the edge

Bullying Signed Network Generation Algorithm
Input: Set of conversation graphs, G_c
Output: Bullying Signed Network $\mathcal{B}$

1. For each conversation graph g_{c_i} in G_c:
   (a) For each set of edges with the same order, sorted ascendingly, compute the bullying score of source node u toward target node v for each edge $e = (u, v)$ as follows:
   $S_{uv} = I_{uv} + ((I_{uv} - S_{vu}) * \alpha)$.
   and then determine the average score of node u for the same set of edges.
   (b) Compute the overall bullying score S of each node in g_{c_i} as follows:
      i. If the node is the *root* node, then:
      $S = \frac{\sum S}{1+2.2(n-1)}$
      ii. Otherwise:
      $S = \frac{\sum S}{2.2(n)}$
2. Construct the bullying signed network graph $\mathcal{B}$ by merging all the conversation graphs together.
3. Return $\mathcal{B}$.

Algorithm 5.2: Bullying Signed Network Generation

responds. In order 2, there are two edges from $P3$ to $P1$ and $P3$ to $P2$ and the bullying score $S_{31} = -61$ and $S_{32} = -61$ is same as I_{31} and I_{32} respectively. The order 3 also has two edges, $P2$ to $P3$ and $P2$ to $P1$. Since the edge $P2$ to $P3$ is a reply to the edge $P3$ to $P2$, the bullying score is calculated as $S_{23} = I_{23} + ((I_{23} - S_{32}) * \alpha) = 0.99$ where $\alpha = 0.6$ was determined by the experiment. Next, the average of the score for the same order of the user is computed i.e., order 2 of user $P3$ is -0.61 and order 3 of user $P2$ is 0.69. Following a similar approach, the bully score S is calculated in Table 5.2 for the second conversation graph g_{c_2} in Figure 5.5b. □

Table 5.1: Bullying score table for g_{c_1}

Tweet #	P1	P2	P3
1		0 (P1)	
2			-0.61 (P1,P2)
3		0.99 (P3) 0.39 (P1)	
Total		0.69	-0.61

Table 5.2: Bullying score table for g_{c_2}

Tweet #	P1	P4	P5	P3
1		-0.18 (P1,P3,P5)		
2			-0.56(P4) -0.42(P1,P3)	
3		0.84(P5) 0.32(P1,P3)		
4				-0.60(P4) -0.16(P5) -0.26(P1)
Total		0.4	-0.49	-0.34

In Step 1b, the bullying score which was computed in the previous step for the users in every conversation graph g_{c_i1} is normalized in $[-1, 1]$. The normalization is performed in two ways i.e., for the user that initiated the conversation, known as root nodes, and the users that are involved in the conversation. For the first type of users, the normalization is computed as $\sum S/(1 + 2.2(n - 1))$ and for the second type of users as $\sum S/2.2(n)$ where, n is the number of times the user occurs in the order and the value 2.2 is computed using $1 + (Maxdiff)(\alpha)$ in which $Maxdiff$ is the range i.e., 2. This normalized score of the users becomes the edge weight to the other users in g_{c_i}.

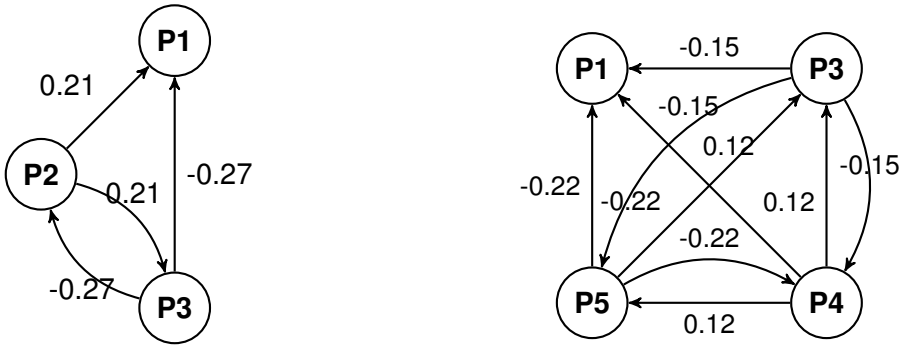


Figure 5.6: Normalized conversation graphs (a) g_{c_1} (b) g_{c_2}

Example 5.2.2. Figure 5.6a and 5.6b shows the normalized conversation graphs for Fig-

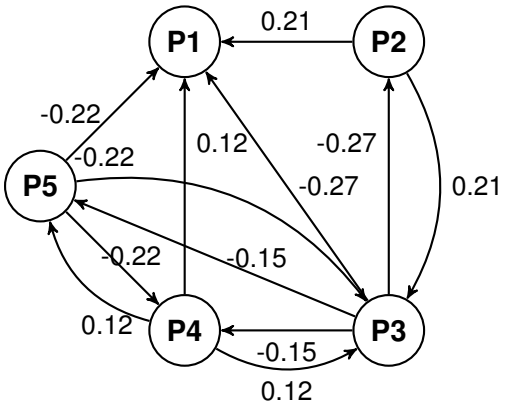


Figure 5.7: Bullying Signed Network

ure 5.5a and 5.5b. The node $P2$ and $P4$ are root nodes with $n = 2$ and nodes $P3$ and $P5$ with n as 1. □

In Step 2, the bullying signed network graph $\mathcal{B}$ is constructed by merging all the conversation graphs G_c. If there is more than one edge i.e., $e = (u, v)$ then a single edge weight is calculated by taking the difference between average and standard deviation of all w_{uv}. Step 4 outputs the bullying signed network graph $\mathcal{B}$.

Example 5.2.3. Figure 5.7 illustrates the bullying signed network by merging the two normalized conversation graphs in Figure 5.6a and 5.6b. From Figure 5.6, it can be seen that there are two different edges from the user $P3$ to $P1$ (-0.34 and -0.12). So, the difference between the average and the standard deviation of the two edges are calculated as -0.57 which is the final edge weight of $P3$ to $P1$ in the bullying signed network. □

5.3 Algorithm 3 - Bully Finding

Given a BSN, with a graph $G_s = (V, E, W)$, where V is the set of users as nodes and E is the set of edges directed from node i to node j, has weight $w_{ij} \in W$ within

the range [-1,1]. Our research is to identify bullies from $\mathcal{B}$ using centrality measure. Centrality is a measure in a network that is used to identify the most important vertices and also to determine how one vertex affects others in a network. The importance of a vertex or node is determined by how high the score is within a network and also defined by the type of the network. Since this research is about social networks the importance is defined as the behaviour. Among several centrality measures, we consider Bias and Deserve (BAD) by Mishra and Bhattacharya [41] because, their measure is computed on how the outgoing edge from a node/user depends on the incoming edges from other nodes/users. However, BAD is modelled on a trust based network i.e., the users that have a propensity to trust/distrust other users. Also, the edge weight denotes trust score rather than the bullying score as in this research.

So, we proposed a centrality measure $A\&M$ Attitude and Merit, similar to that of BAD to identify bullies from our proposed signed network $\mathcal{B}$. Merit is a measure of the opinion (good or bad) that the other nodes have towards a particular node and Attitude is a measure of the behaviour of a node towards the other node. However, in a given bullying singed network, the attitude or likes or dislike of a node towards other nodes in the network is not known. Therefore the expressions to compute the Merit and Attitude metrics in a mutually recursive manner are:

$$M^{n+1}(j) = \frac{1}{2|in(j)|} \sum_{k \in in(j)} (w_{kj})(A^n(k)) \tag{5.1}$$

$$A^{n+1}(i) = \frac{1}{2|out(i)|} \sum_{j \in out(i)} (w_{ij} + X_{ij}) \tag{5.2}$$

Let $in(j)$ denote the set of all incoming edges to node j and $out(i)$ denote the set of all outgoing edges from node i. Normalization is done to maintain the value in the range

of [-1, 1]. An auxiliary variable X_{ij} is introduced to measure the effect of the merit score of a node j on its incoming edge to node i. Since merit is about whether the node is considered to be good or bad, it is calculated by the sum of all its incoming edges from other nodes. Likewise, since attitude is about the particular node's view of others, it is calculated using the outgoing edges of a node towards others and its corresponding merit score in the network. Although we use two metrics similar to BAD, the calculation of the incoming and the outgoing edges of a node differs. Since bias in BAD is about how truthfully it rates other nodes, it is calculated by the difference in the edge weight and the true trust of a node(deserve). The explanation of the proposed metric follows.

$$X_{ij} = \begin{cases} M(j) & \text{if } (w_{ij} \times M(j)) > 0 \\ -M(j) & \text{otherwise} \end{cases}$$

From the above expression, it can be seen that if the outgoing edge weight from node i to node j has a positive value and the merit score of node j is also positive, then the attitude of node i to j is calculated by the sum of both values. If the outgoing edge weight from node i to j is negative and the merit score of node j is positive or vice-versa, then the attitude of node i to j is calculated by subtracting the merit score from the edge weight. This means if a node has a positive edge weight towards a benign merit node then the attitude score increases. Similarly, holding a negative edge weight towards a benign merit node decreases that node's attitude score. However, if a node has a positive edge weight towards a negative merit node, the attitude of a node decreases.

From Eq. 5.1 and Eq. 5.2, the attitude of a node depends on the merit of its neighbours and vice versa. A fixed-point iteration method is used to obtain the solution. The Merit and Attitude of node i at iteration n are denoted by $A^n(i)$ and $M^n(i)$ respectively. The

Bully Finding Algorithm

Input: Bullying Signed Network $G_s = (V, E, W)$
Output: List of bullies and its attitude score $L = [(u_1, s_1), (u_2, s_2), \ldots, (u_{|L|}, s_{|L|})]$

1. Initialize $M^0(v) = -1$ and $A^0(v) = -1, \forall v \in V$.
2. Set iteration index i = 1
   (a) For each $v \in V$ compute merit score $M^i(v) = \frac{1}{2|in(v)|} \sum_{u \in in(v)} (w_{uv})(A^{i-1}(u))$ where $|in(v)|$ is the number of incoming edges to the node v
   (b) For each $u \in V$ compute attitude score $A^i(u) = \frac{1}{2|out(u)|} \sum_{v \in out(u)} (w_{uv} + X_{uv})$ where $|out(u)|$ is the number of outgoing edges from the node u
3. If there exist atleast one $v \in V : M^i(v) \neq M^{i-1}(v)$ or $A^i(v) \neq A^{i-1}(v)$
   (a) Increase the iteration index $i = i + 1$
   (b) Repeat step 2a & 2b for each iteration
4. For each $v \in V$ add the node and its corresponding attitude score, whose score value greater than 0 to the list L
5. Return L

Algorithm 5.3: Bully Finding Algorithm

proposed algorithm 5.3 is designed to compute merit and attitude scores for each node in the network. Initially, we start with an Merit and Attitude score of -1 (i.e, the first iteration) in step 1. In step 2a, the merit scores for each node are updated using the attitude scores from the previous iteration. In step 2b, the attitude scores are updated using the newly updated Merit scores in the same iteration. Both Merit and Attitude scores are mutually recursive and are updated till both the scores converges in step 3. The scores of Merit and

Attitude from the last iteration are the final scores. In the final step 4, all the nodes whose attitude score is less than zero are added the list L along with the user's attitude score.

Table 5.3: Example showing the values of the graph (Figure 5.7) after each iteration. (A denotes attitude and M denotes merit)

No.	P1		P2		P3		P4		P5	
	M	**A**	**M**	**A**	**M**	**A**	**M**	**A**	**M**	**A**
0	-1	-	-1	-1	-1	-1	-1	-1	-1	-1
1	0.02	-	0.01	0.11	-0.01	-0.13	0.09	0.06	0.01	-0.13
2	0.01	-	0.02	0.11	0.1	-0.11	0.01	0.06	0.0	-0.11
3	0.01	-	0.01	0.11	0.00	-0.11	0.01	0.06	0.0	-0.11
4	0.01	-	0.01	0.11	0.00	-0.11	0.01	0.06	0.0	-0.11

Table 5.4: Final Attitude & Merit values

	P1	P2	P3	P4	P5
Merit	0.01	0.01	0.0	0.01	0.0
Attitude	-	0.11	-0.11	0.06	-0.11

Example 5.3.1. Table 5.3 demonstrates the value of Attitude and Merit that are updated after each iteration by applying algorithm 5.3 to Figure 5.7. The Aattitude column of the node $P1$ is blank because there are no outgoing edges from $P1$. The Table 5.4 shows the final attitude and merit score of the nodes. It can be seen that, the node $P3$ and $P5$ are a bully with the confidence score of 0.11 and 0.11, respectively. □

CHAPTER 6

ALGORITHM ANALYSIS

In chapter 6, we show the proof of convergence of the centrality measure and perform a complexity analysis of our proposed approach.

6.1 Convergence of Centrality Measure

We start the convergence proof by showing that the difference between the Attitude of a node at any iteration and the infinite iteration is bounded, which then leads to convergence by proving the error bound $\in, << 1$. After a certain iteration t, the Attitude score of that iteration becomes close to A^{∞}. Since Merit of a node can be expressed in terms of the Attitude of other nodes, this implies that Merit values exhibit similar properties.

Proposition 6.1.1. *The Attitude and Merit(A&M) of a node at any iteration n converges to the infinite iteration by a constant.*

Proof : By using mathematical induction we prove the convergence of attitude. Given its definition, the attitude score $A^{\infty}(i)$ and $A^{n+1}(i)$ can be written as,

$$A^{\infty}(i) = \left| \frac{1}{2|out(i)|} \sum_{j \in out(i)} \left\{ w_{ij} \pm \frac{1}{2|in(j)|} \sum_{k \in in(j)} (w_{kj} \times A^{\infty}(k)) \right\} \right|$$

$$A^{n+1}(i) = \left| \frac{1}{2|out(i)|} \sum_{j \in out(i)} \left\{ w_{ij} \pm \frac{1}{2|in(j)|} \sum_{k \in in(j)} (w_{kj} \times A^{n}(k)) \right\} \right|$$

Base case: For n=1, we have

$$
\begin{aligned}
&|A^{\infty}(i) - A^{1}(i)| \\
&\quad = \left|\tfrac{1}{2|out(i)|}\sum_{j\in out(i)}\left\{w_{ij} \pm \tfrac{1}{2|in(j)|}\sum_{k\in in(j)}(w_{kj}\times A^{\infty}(k))\right\}\right| - \\
&\qquad \left|\tfrac{1}{2|out(i)|}\sum_{j\in out(i)}\left\{w_{ij} \pm \tfrac{1}{2|in(j)|}\sum_{k\in in(j)}(w_{kj}\times A^{0}(k))\right\}\right| \\
&\quad = \left|\tfrac{1}{2|out(i)|}\sum_{j\in out(i)}\left\{w_{ij} \pm \tfrac{1}{2|in(j)|}\sum_{k\in in(j)} w_{kj}(A^{\infty}(k)) - A^{0}(k)\right\}\right| \\
&\quad \leq \tfrac{1}{2|out(i)|}\sum_{j\in out(i)}\left\{|w_{ij}| \pm \tfrac{1}{2|in(j)|}\sum_{k\in in(j)}|w_{kj}||(A^{\infty}(k)) - A^{0}(k)|\right\} \\
&\qquad\qquad [\because |x.y| \leq |x||y|] \\
&\quad \leq \tfrac{1}{2|out(i)|}\sum_{j\in out(i)}\left\{\tfrac{1}{2|in(j)|}\sum_{k\in in(j)}|(A^{\infty}(k)) - A^{0}(k)|\right\} \\
&\qquad\qquad [\because |w_{ij}|\ and\ |w_{kj}| \leq 1] \\
&\quad \leq \tfrac{1}{2|out(i)|}\sum_{j\in out(i)}\left\{\tfrac{1}{2|in(j)|}\sum_{k\in in(j)} 2\right\}
\end{aligned}
$$

Since $A(k) \in [-1, +1]$, we have $|A^{\infty}(k) - A^{0}(k)| \leq 2$

$$
\begin{aligned}
&\quad \leq \tfrac{1}{2|out(i)|}\sum_{j\in out(i)}\left\{\tfrac{1}{2|in(j)|}2|in(j)|\right\} \\
&\quad = \tfrac{1}{2}
\end{aligned}
$$

Induction step : We assume the bound to be true for $A^{n}(i)$ so, by the hypothesis $|A^{\infty}(i) - A^{n}(i)| \leq \frac{1}{2^{n}}$. In the $(n+1)^{th}$ iteration,

$$
\begin{aligned}
&|A^{\infty}(i) - A^{n+1}(i)| \\
&\quad = \left| \tfrac{1}{2|out(i)|} \textstyle\sum_{j \in out(i)} \left\{ w_{ij} \pm \tfrac{1}{2|in(j)|} \textstyle\sum_{k \in in(j)} w_{kj}(A^{\infty}(k)) - A^{n}(k) \right\} \right| \\
&\quad \leq \tfrac{1}{2|out(i)|} \textstyle\sum_{j \in out(i)} \left\{ \tfrac{1}{2|in(j)|} \textstyle\sum_{k \in in(j)} |(A^{\infty}(k)) - A^{n}(k)| \right\} \\
&\quad \leq \tfrac{1}{2|out(i)|} \textstyle\sum_{j \in out(i)} \left\{ \tfrac{1}{2|in(j)|} \textstyle\sum_{k \in in(j)} \tfrac{2}{2^n} \right\} \\
&\quad = \tfrac{1}{2^{n+1}}
\end{aligned}
$$

Therefore, the error is bounded by an inverse exponential function. Thus, we conclude that a convergence has been achieved in determining the measures 'Attitude' and 'Merit'. □

6.2 Complexity Analysis

Proposition 6.2.1. *The overall complexity of the proposed approach in the average case is* $\mathcal{O}((k \times l + \log n) \times n)$

Proof : We can determine the time complexity of the proposed approach in three phases: constructing conversation graph, constructing bullying signed network and bully finding.

Constructing conversation graphs phase. In the constructing conversation phase, the runtime complexity is the time taken to construct m conversations from n tweets and then to generate graphs from the constructed conversations.

Initial sorting of tweets uses merge sort, which takes a computational time of $\mathcal{O}(n \log n)$. The conversation is constructed by doing a binary search on the $\mathcal{DID}$ and $\mathcal{SID}$ of the

conversation tweet and the current tweet respectively, leading to m conversations with a computational time of $\mathcal{O}(n \log n)$. The cost for generating the graph from the conversations is $\mathcal{O}(m)$. Therefore the average computational cost to construct the conversation graphs is $\mathcal{O}(n \log n + n \log n + m) = \mathcal{O}(n \log n + m)$

Constructing bullying signed network phase. In the constructing the bullying signed network phase, we traverse though each conversation graph where the bullying score is calculated for each node with respect to the edges with same order. For each conversation graph m, the maximum number of nodes in the worst case is k. Therefore the total computational cost is $\mathcal{O}(n.k + m.k)$

Bully finding phase. In the bully finding phase, the runtime is the time taken to detect the bullying users using an Attitude and Merit centrality. For each l number of iteration, the A&M centrality touches each edge at most twice. Therefore, the average case in detecting bullies in each iteration is $\mathcal{O}(2n.k)$ and for the given l iteration it is $\mathcal{O}(n.k.l)$

Therefore, the overall complexity of the proposed approach in the average case is:

$\mathcal{O}((k.l + \log n)n + k.m)$ = $\mathcal{O}(k.l + \log n)n$ since $m, k << n$.

CHAPTER 7

EXPERIMENTAL EVALUATION

In this chapter, we evaluate the performance of the proposed algorithms. First, we present the data used in our evaluation, second we discuss the implementation details, and the way we process it to build ground truth. Finally, we present the experimental results which include determining the coefficients α, β and γ, utility and scalability.

7.1 Dataset

In this paper, we rely on Twitter's Streaming API, which provides free access to 1% of all tweets. The API returns each tweet in a JSON format, with the content of the tweet, some metadata (e.g., creation time, source ID, destination ID, reply/retweet, etc.) as well as information about the poster (e.g., username, followers, friends) with an total of 5.6M tweets within a six month time-frame. We then extract features like users, text, replyuser, mentions and network based features like source ID, destination ID from the Twitter JSON.

7.2 Implementation and Setup

We implemented our algorithm in Java, and our experiments were conducted on a machine equipped with an Intel(R) Core(TM) i7-8550U CPU @ 2.00GHz processor and 16.0 GB RAM, running Windows 10 64-bit operating system.

We employed Amazon *Mechanical Turk* (mturk) workers to respond to an online survey that we developed. We provided 2700 surveys with each survey consisting of 10 conversations. Each survey was assigned to three workers to classify the bullying behavior of the users in the conversations according to predefined labels (strongly positive, likely positive, likely negative and strongly negative). Overall, the workers rated 27000 conversations, which were extracted from the set of raw Twitter data by using algorithm 5.1. The MTurk UI enables requesters to create and publish HITs in a batch when processing many HITs of the same type thus saving time. For our study, we created a csv file that contained 2700 HITs. MTurk automatically created a separate HIT for each set of conversation in the csv file. The results to rate each users involved in the set of conversations are obtained from the workers. There was not marked variation in rating provided by the workers. Finally, the results are combined for the users to form the ground truth.

7.3 Determining optimal values for coefficients α, β and γ

Recall that $I_{uv} = \beta * SA + \gamma * CS$ in algorithm 5.1 and $S_{uv} = I_{uv} + ((I_{uv} - S_{vu}) * \alpha)$ in algorithm 5.2. To determine the coefficient β and γ for bullying indicator I and α for the bullying score S, we randomly generate input tweets of varying length and experiment for different values of α, β and γ. After experimenting with different values, we found that the coefficient values of $\beta \geq 0.6$, $\gamma \leq 0.4$ and $\alpha \leq 0.6$ to provide the greatest accuracy. The accuracy was measured with $\beta \geq 0.6$ and $\gamma \leq 0.4$ for every $\alpha \leq 0.6$ with respect to the ground truth, using the F1 Measure [48].

Figure 7.1 depicts the optimal values for the coefficients α, β and γ with respect to the β and γ values, which are set from 60 to 90 and 40 to 10 respectively. We use three different α values for every bullying indicator coefficients β and γ, which varies from 0.4 to 0.6.

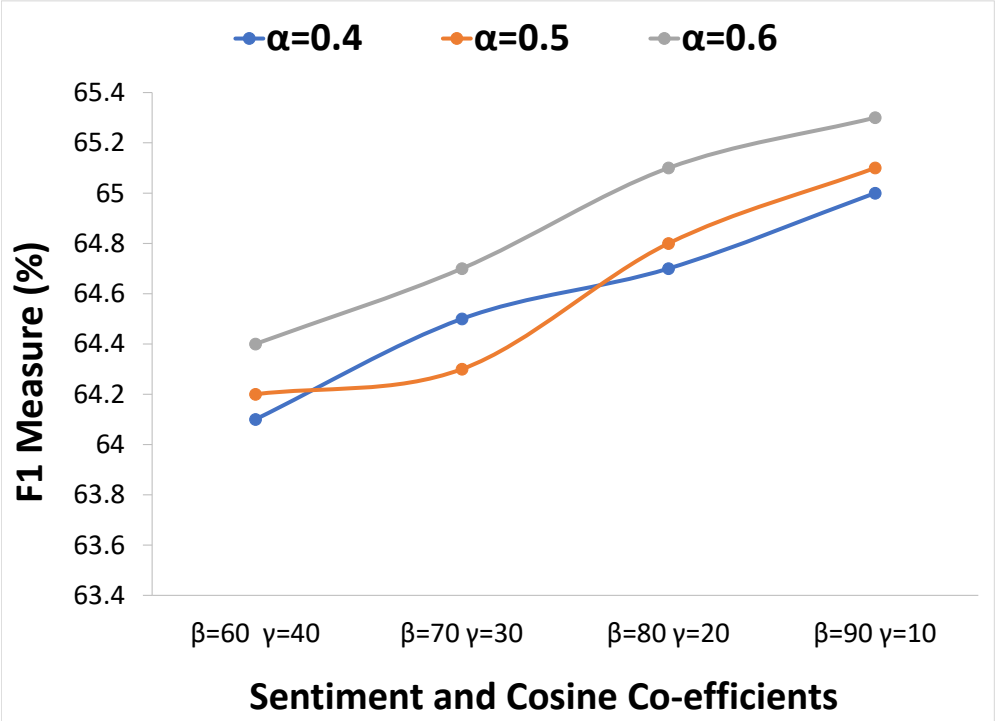


Figure 7.1: Optimal values for coefficients α, β and γ

In our approach, we observe that the F1 measure increases linearly when the coefficients β increases and γ decreases. We also observe that when we increase the α value, the F1 measure increases in all the cases. This indicates that the sentiment analysis has more impact on the bullying indicator than the cosine similarity. Similarly, the response to a tweet has a direct effect on the bullying score.

7.4 Utility

We briefly introduce our evaluation metrics that will be used to determine the accuracy of our approach.

- *AccuracyCM [40]*

  The accuracy measure is the ratio of number of bully users detected to the total number of bullies. It does not perform well with imbalanced data sets.

$$AccuracyCM = \frac{\#\ of\ detected\ bullies}{total\ number\ of\ bullies}$$

- *Precision and Recall [45]*

  Precision and Recall are evaluation metrics used in binary classification tasks. Precision is the measure of exactness and recall is the measure of completeness. They are defined as follows:

$$Precision = \frac{\#\ of\ true\ bullies\ detected}{total\ number\ of\ detected\ users}$$

$$Recall = \frac{\#\ of\ true\ bullies\ detected}{total\ number\ of\ true\ bullies}$$

  In simple terms, high precision means that an algorithm returned substantially more bully users, while high recall means that an algorithm returned most of the bullies.

- *F1 Measure [48]*

  F1 Measure is the Harmonic Mean between precision and recall. The range for F1 is [0, 1]. It measures how many bullies are identified correctly and how robust it is. Mathematically, it can be expressed as :

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1 Measure attempts to find a balance between precision and recall. The greater the F1 Measure, the better is the performance of our approach.

To determine the accuracy of our proposed centrality measure, Attitude and Merit, we compare all the evaluation metrics discussed above with respect to the number of users increasing linearly from 500 to 1700 users. Figure 7.2 illustrates the accuracy utility to measure the metrics (accuracyCM, precision, recall and F1 Measure) with respect to the number of users generated from the algorithm 5.2 as the input.

We observed that the AccuracyCM metric is about 60% and can be biased in the case of unbalanced datasets, however it produces better results when false positives (is an error in bully detection in which a detection result improperly indicates that a user is bully, when in reality the user is not a bully) and false negatives (is an error in which a test detection improperly indicates that a user is not bully, when in reality the user is a bully.) are almost even. In the case of uneven distribution of data, we use F1 Measure, which is at 70% while the precision and recall are consistently at 77% and 65% respectively. Therefore from the Figure 7.2 it can be seen that, the precision out perform other metrics i.e., higher the precision means our algorithm identifies more bullies precisely among the total number of bully users. The percentages mentioned above for all the metrics remained almost consistent even with increase in the number of users.

Next, we compare the performance of our proposed centrality measure Attitude and Merit with the research work done by Mishra and Bhattacharya [41] - Bias and Deserve. We compared the F1-score in term of accuracy achieved with respect to the number of users generated from the algorithm 5.2 as the input. Figure 7.3 elucidates the comparison of the centrality measures w.r.t. the number of users increasing linearly from 500 to 1700 users.

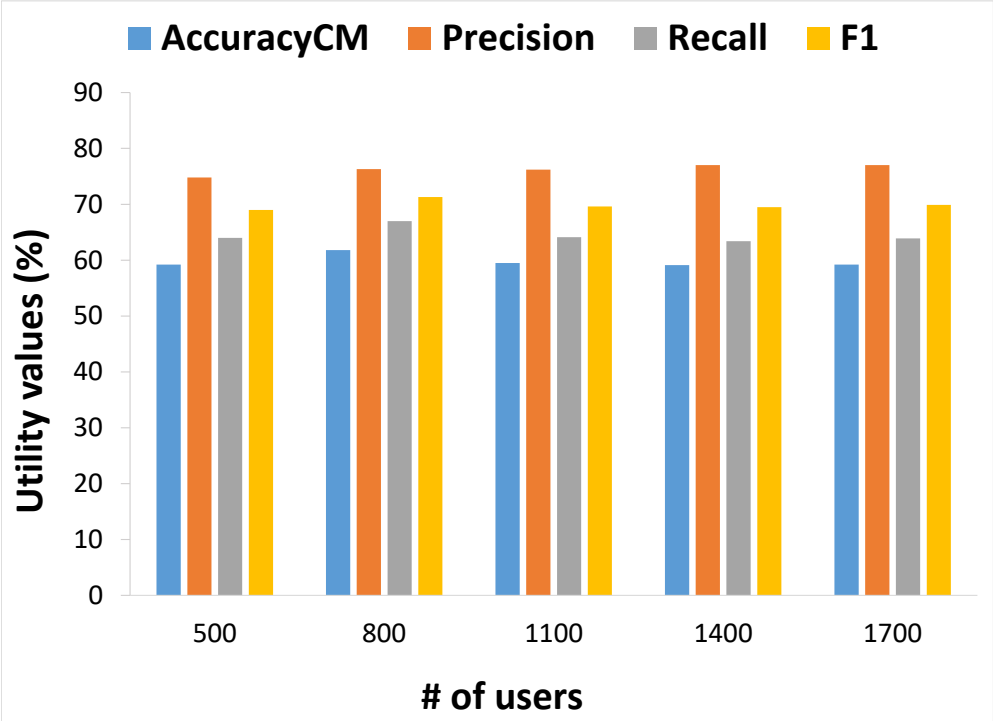


Figure 7.2: Accuracy with respect to the number of users

In our approach, we observed that $A\&M$ has an accuracy of 70%. Also, our centrality measures outperform BAD in all the cases i.e., number of users. The accuracy of Bias and Deserve gets decreased as number of users increases whereas the proposed centrality measures Attitude and Merit stays consistent. This can have multiple reasons behind it. First of all, the bias score of a node with highly positive bias decreases when it has outgoing edge with positive weight whereas in $A\&M$, the Attitude score increases when positive node has outgoing edge with positive weight. Next, when calculating the deserve for a node, the bias value is taken in range of $[0, 1]$ whereas in $A\&M$, merit is calculated with

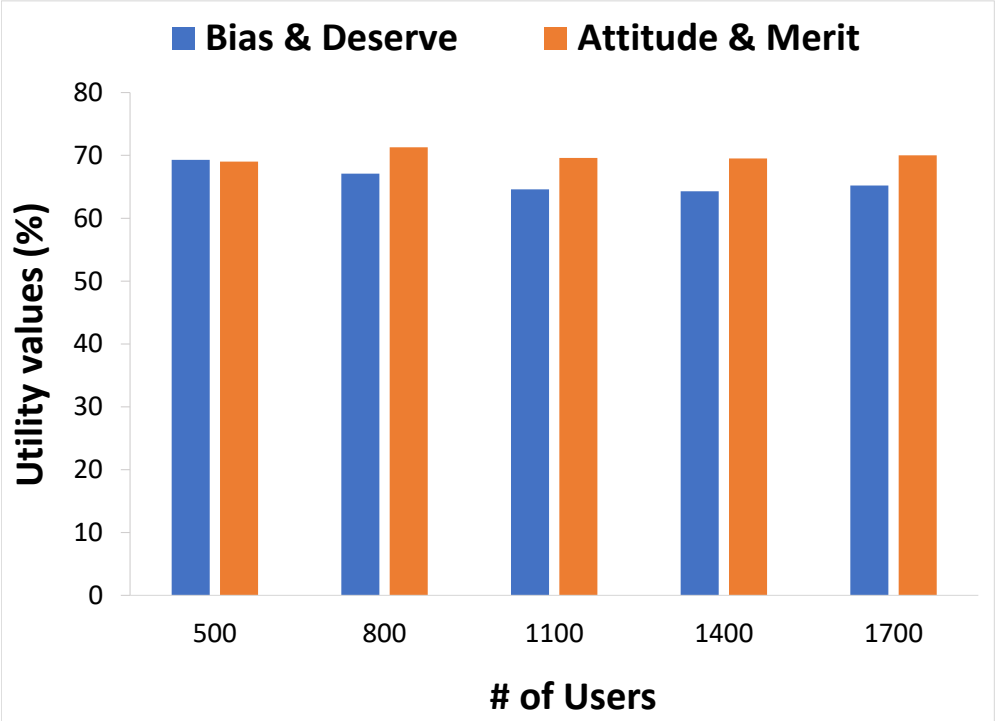


Figure 7.3: Comparative evaluation of the proposed centrality measure Attitude and Merit with Bias and Deserve

the attitude value in range of $[-1, 1]$. Furthermore, BAD does not perform well when a node has fewer outgoing and incoming edges. Nevertheless it is still outperformed by the $A\&M$ centrality.

7.5 Scalability

We measure the scalability of BullyNet with respect to the number of tweets and observe the run-times of our three algorithms: conversation graphs generation, bullying

signed network generation and bully finding with optimal values for coefficients α, β and γ set at 0.6, 0.9 and 0.1 respectively.

We observed that running a dataset with 1M records takes up to 8 min for the BullyNet algorithm and the runtime increases linearly as the record size increases linearly from 1M to 5M. Figure 7.4 depicts the runtime for the records size from 1M to 5M for each dataset. We also observed that the most dominant algorithm of our experiment is conversation graphs generation which took the majority of run time i.e approximately 70% of total execution time of the three algorithms. This is due to the fact that the conversation graphs have to calculate sentiment analysis and cosine similarity for each tweet and then calculate bullying indicator I as the edge weight for each conversation graph.

We observed that there is a linear increase in total runtime with increase in number of tweets. However, we also observed that the bullying signed network generation algorithm runtime didn't grow linearly with the increase in records, rather it tends to remain constant. This is because, there are k number of nodes in m conversation graphs. So, to calculate the bullying score for each graphs it takes $\mathcal{O}(k)$. Hence, it does not affect the run-time with the growth of number of tweets.

We can observe that similar to the first algorithm, the runtime of the third algorithm also increases linearly with record size. The variation is attributed to the increase in the number of users in each tweet resulting in corresponding increase in computation time for centrality measures.

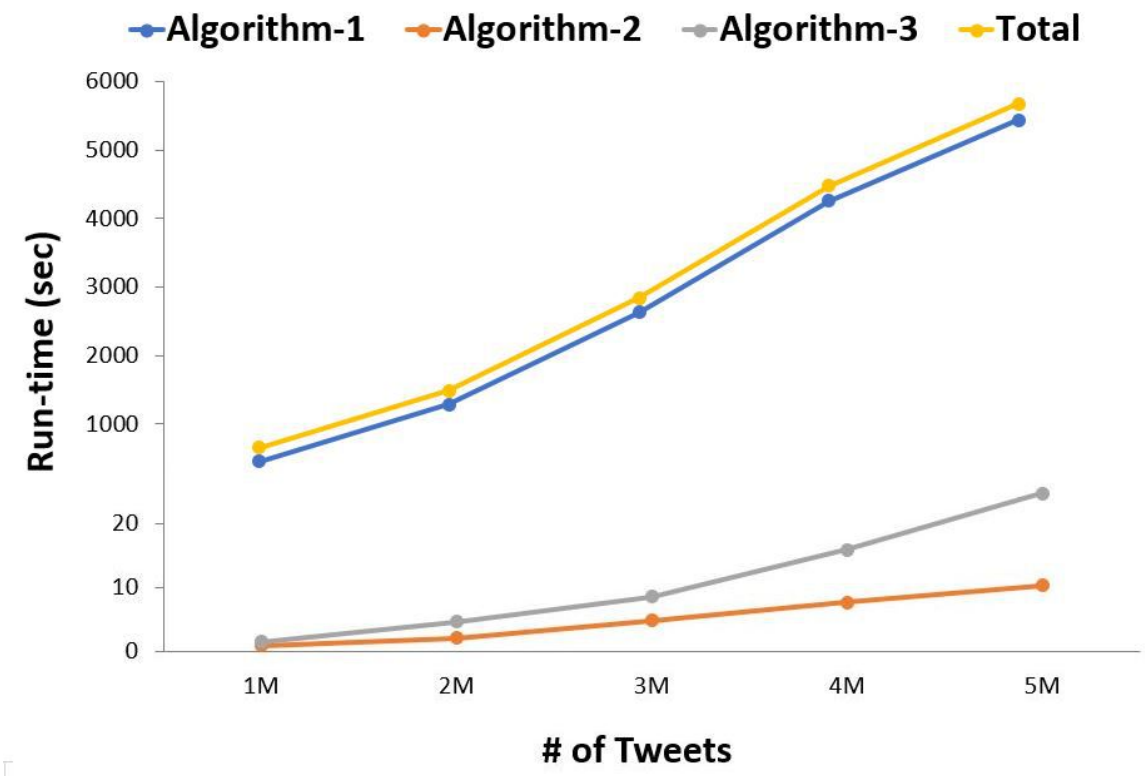


Figure 7.4: Scalability with respect to the number of tweets

CHAPTER 8

CONCLUSION AND FUTURE WORK

8.1 Summary

Although the digital revolution and the rise of social media enabled great advances in communication platforms and social interactions, a wider proliferation of harmful behavior known as bullying has also emerged. Aiming to address this bullying, this thesis presents a novel framework to identify bully users from the Twitter social network. We performed extensive research on mining signed networks for better understanding of the relationships between users in social media, to build a signed network (SN) based on bullying tendencies. We observed that by constructing conversations based on the context as well as content, we can effectively identify the emotions and the behavior behind bullying. In our experimental study, the evaluation of our proposed centrality measures to detect bullies from signed network, we achieved 70% accuracy with 77% precision in identifying bullies.

In chapter 2, we presented and discussed the building blocks of our thesis that are very important for the implementation of this work. We explain the sentiment analysis, the different techniques to analyze the sentiment of the message, the cosine similarity, the centrality measures and the different types of measures used in signed networks.

In chapter 3, we examined the related work done in the field of cyberbullying detection. We did extensive research on a signed network focusing on node classification, balance theory and measure designed to analyze the signed network. We made a comparative

evaluation and presented in Table 3.1 for cyberbullies detection, signed networks and centrality measures techniques.

In chapter 4, we formulate and present the output our problem.

In chapter 5, we propose our solution for building a BullyNet three phase algorithm and explain the execution of all the phases. The proposed solution achieves a high accuracy and is scalable with large datasets.

In chapter 6, we provided proof for convergence of our proposed centrality measures and the complexity analysis of the proposed approach.

In chapter 7, we analyzed the performance of our approach. We carried out experiments with ground truth to establish the accuracy and scalability of our solution. The experimental results show that our approach achieves high accuracy, is scalable, and is precise in detecting bullies from the dataset.

Overall, the objective of this thesis work is to design and implement an efficient and scalable approach for identifying bullies on the Twitter network with high accuracy.

8.2 Future Work

In conclusion, there are several open questions that deserve further investigation. First, our approach focuses on extracting emotions and behavior from texts and emojis in tweets. However, it would be interesting to investigate images and videos, given that many users use them to bully others. Second, it does not distinguish between bully and aggressive users. Devising new algorithms or techniques to distinguish bullies from aggressors would prove critical in better identification of cyberbullies. Another topic of interest would be to study the relationship between conversation graph dynamics and geographic location

and how these dynamics are affected by the geographic dispersion of the users? Are the proximity increase the bullying behaviour?

REFERENCES

[1] Apoorv Agarwal, Fadi Biadsy, and Kathleen R Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the Conference of the European Chapter of the ACL*, pages 24–32, 2009.

[2] Smriti Bhagat, Graham Cormode, and S Muthukrishnan. Node classification in social networks. In *Social network data analytics*, pages 115–148. 2011.

[3] Petko Bogdanov, Nicholas D Larusso, and Ambuj Singh. Towards community discovery in signed collaborative interaction networks. In *Proceedings of the IEEE International Conference on DMW*, pages 288–295, 2010.

[4] Phillip Bonacich and Paulette Lloyd. Calculating status with negative relations. *Social networks*, 26(4):331–338, 2004.

[5] Piotr Borzymek and Marcin Sydow. Trust and distrust prediction in social network with combined graphical and review-based attributes. In *Proceedings of the KES-AMSTA*, pages 122–131, 2010.

[6] Ulrik Brandes and Dorothea Wagner. Analysis and visualization of social networks. In *Graph drawing software*, pages 321–340. 2004.

[7] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[8] Erin E. Buckels, Paul D. Trapnell, and Delroy L. Paulhus. *Trolls just want to have fun*, pages 67:97–102. 2014.

[9] Cyberbullying Research Center. https://cyberbullying.org/bullying-laws.

[10] 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 ACM on WebSci*, pages 13–22, 2017.

[11] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the PASSAT and SCSM*, pages 71–80, 2012.

[12] Kai-Yang Chiang, Nagarajan Natarajan, Ambuj Tewari, and Inderjit S Dhillon. Exploiting longer cycles for link prediction in signed networks. In *Proceedings of the ACM international CIKM*, pages 1157–1162, 2011.

[13] Council Of Europe children's rights. https://www.coe.int/en/web/children/bullying.

[14] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *Proceedings of the European Conference on IR*, pages 693–696, 2013.

[15] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *In Proceedings of the ACM TiiS*, 2(3):18, 2012.

[16] Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *Proceedings of the international AAAI WSM*, 2011.

[17] Thomas DuBois, Jennifer Golbeck, and Aravind Srinivasan. Predicting trust and distrust in social networks. In *Proceedings of the IEEE international PASSAT conference on SC*, pages 418–424, 2011.

[18] Patxi Galn-Garca, J.G. De La Puerta, C.L. Gmez, Igor Santos, and Pablo Bringas. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. 24:42–53, 2014.

[19] Lise Getoor and Christopher P Diehl. Link mining: a survey. *In Proceedings of the ACM SIGKDD Explorations Newsletter*, 7(2):3–12, 2005.

[20] Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the international conference on WWW*, pages 403–412, 2004.

[21] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. 2011.

[22] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the international conference on WWW*, pages 517–526, 2002.

[23] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. *In Proceedings of the CoRR*, abs/1503.03909, 2015.

[24] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD international conference on KDD*, pages 168–177, 2004.

[25] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Social spammer detection with sentiment information. *In Proceedings of IEEE ICDM*, pages 180–189, 2014.

[26] Clayton J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI WSM*, 2014.

[27] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the international conference on WWW*, pages 640–651, 2003.

[28] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ACM SIGKDD international conference on KDD*, pages 137–146, 2003.

[29] Cristobald de Kerchove and Paul Van Dooren. The pagetrust algorithm: How to rank web pages when negative links are allowed? In *Proceedings of the SIAM International Conference on Data Mining*, pages 346–352, 2008.

[30] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the international conference on CL*, page 1367, 2004.

[31] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *In Proceedings of the ACM (JACM)*, 46(5):604–632, 1999.

[32] Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. Accurately detecting trolls in slashdot zoo via decluttering. In *Proceedings of IEEE/ACM ASONAM*, pages 188–195, 2014.

[33] Jérôme Kunegis, Julia Preusse, and Felix Schwagereit. What is the added value of negative links in online social networks? In *Proceedings of the International Conference on WWW*, pages 727–736, 2013.

[34] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the international conference on WWW*, pages 641–650, 2010.

[35] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI CHI*, pages 1361–1370, 2010.

[36] Yadong Li, Jing Liu, and Chenlong Liu. A comparative analysis of evolutionary and memetic algorithms for community detection from signed social networks. *Soft Computing*, 18(2):329–348, 2014.

[37] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Proceedings of the ASIS&T*, 58(7):1019–1031, 2007.

[38] Bo Long, Zhongfei Mark Zhang, Xiaoyun Wu, and Philip S Yu. Spectral clustering for multi-type relational data. In *Proceedings of the ICML*, pages 585–592, 2006.

[39] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Proceedings of the Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[40] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, pages 283–298, 1978.

[41] Abhinav Mishra and Arnab Bhattacharya. Finding the bias and prestige of nodes in networks based on trust scores. *In Proceedings of the international conference on WWW*, 2011.

[42] Le Falher Ortega, José A Troyano, FermíN L Cruz, Carlos G Vallejo, and Fernando EnríQuez. Propagation of trust and distrust for the detection of trolls in a social network. *Computer Networks*, 56(12):2884–2895, 2012.

[43] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the annual meeting on Association for Computational Linguistics*, page 271, 2004.

[44] Justin W. Patchin and Sameer Hinduja. 2016 cyberbullying data, 2017.

[45] James W Perry, Kent Allen, and Madeline M Berry. Machine literature searching x. machine language; factors underlying its design and development. *American Documentation (pre-1986)*, page 242, 1955.

[46] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. 1980.

[47] Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In *Proceedings of the International SW conference*, pages 351–368, 2003.

[48] Yutaka Sasaki et al. The truth of the f-measure. *Teach Tutor mater*, pages 1–5, 2007.

[49] Moshen Shahriari and Mahdi Jalili. Ranking nodes in signed social networks. *Social network analysis and mining*, 4(1):172, 2014.

[50] Vivek K. Singh, Qianjia Huang, and Pradeep K. Atrey. Cyberbullying detection using probabilistic socio-textual information fusion. *In Proceedings of the IEEE/ACM ASONAM*, pages 884–887, 2016.

[51] Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Y Liu, and Christopher H Griffin. Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the IEEE/ACM ASONAM*, pages 280–285, 2015.

[52] Panagiotis Symeonidis and Nikolaos Mantas. Spectral clustering for link prediction in social networks with positive and negative links. *Social Network Analysis and Mining*, 3(4):1433–1447, 2013.

[53] Jiliang Tang, Charu Aggarwal, and Huan Liu. Node classification in signed social networks. In *Proceedings of the SIAM ICDM*, pages 54–62, 2016.

[54] Jiliang Tang, Charu Aggarwal, and Huan Liu. Recommendations in signed social networks. In *Proceedings of the International Conference on WWW*, pages 31–40, 2016.

[55] Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. A survey of signed network mining in social media. *In Proceedings of the ACM Comput. Surv.*, 49(3):42:1–42:37, 2016.

[56] Jiliang Tang, Xia Hu, and Huan Liu. Social recommendation: a review. *Social Network Analysis and Mining*, 3(4):1113–1133, 2013.

[57] Lei Tang and Huan Liu. Community detection and mining in social media. *Synthesis lectures on data mining and knowledge discovery*, 2(1):1–137, 2010.

[58] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the annual meeting on association for CL*, pages 417–424, 2002.

[59] Traag Vincent, Nesterov Yurii, and Van Dooren. Paul. Exponential ranking: Taking into account negative links. *In Proceedings of the Social Informatics*, pages 192–202, 2010.

[60] Suhang Wang, Charu Aggarwal, Jiliang Tang, and Huan Liu. Attributed signed network embedding. In *Proceedings of the ACM on CIKM*, pages 137–146, 2017.

[61] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the HLT Conference and Conference on EMNLP*, 2005.

[62] Zhaoming Wu, Charu C. Aggarwal, and Jimeng Sun. The troll-trust model for ranking in signed networks. In *Proceedings of the ACM International Conference on WSDM*, pages 447–456, 2016.

[63] Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. Fast learning for sentiment analysis on bullying. In *Proceedings of the First International WISDOM*, pages 10:1–10:6, 2012.

[64] Tongda Zhang, Haomiao Jiang, Zhouxiao Bao, and Yingfeng Zhang. Characterization and edge sign prediction in signed networks. *In Proceedings of the JIII*, 1(1), 2013.

[65] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the ICDCN*, 2016.

[66] Shenghuo Zhu, Kai Yu, Yun Chi, and Yihong Gong. Combining content and link for classification using matrix factorization. In *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval*, pages 487–494, 2007.

[67] Cai-Nicolas Ziegler and Georg Lausen. Propagation models for trust and distrust in social networks. *Information Systems Frontiers*, 7(4-5):337–358, 2005.