e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703075 |

BullyNet: Unmasking Cyberbullies on Social Networks

B V Srikanth[1], E Vasanth Rao[2], K Lahari[3], J Venugopal[4]

Assistant Professor, Department of Computer Science and Engineering, Anurag University, Ghatkesar, Hyderabad, India [1]

Student, Department of Computer Science and Engineering, Anurag University, Ghatkesar, Hyderabad, India [2]

Student, Department of Computer Science and Engineering, Anurag University, Ghatkesar, Hyderabad, India [3]

Student, Department of Computer Science and Engineering, Anurag University, Ghatkesar, Hyderabad, India [4]

ABSTRACT: 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 paper, 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. We analyze tweets to determine their relation to cyberbullying, while considering the context in which the tweets exist in order to optimize their bullying score. We also propose a centrality measure to detect cyberbullies from a cyberbullying signed network, and we show that it outperforms other existing measures.

We experiment on a dataset of 5.6 million tweets and our results shows that the proposed approach can detect cyberbullies with high accuracy, while being scalable with respect to the number of tweets.

I. 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 Face book, 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 [1], link predictions [2], visualization, and analysis of social networks [3].

While the growth of social media has created an excellent platform for communications and information sharing, it has also created a new platform for malicious activities such as spamming [4], trolling [5], and cyber bullying [6]. According to the Cyber bullying Research Center (CRC) [7], cyber bullying 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, cyber bullying contains hurtful messages which are present online for a long time. These messages can be accessed worldwide, and are often irrevocable. Laws about cyber bullying and how it is handled differ from one place to another.

Popular social media platforms such as Face book and Twitter are very vulnerable to cyber bullying due to the popularity of these social media sites and the anonymity that the internet offers to the perpetrators. Although strict laws exist to punish cyber bullying, there are very less tools available to effectively combat cyber bullying. Social media platforms provide users with the option to self-report abusive behavior and content in addition to providing tools to deal with bullying.

The body of work produced by the research community with regards to cyber bullying in social networks also needs to be expanded to get better insights and help develop effective tools and techniques to tackle the issue.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703075 |

II. LITERATURE SURVEY

BullyNet: Unmasking Cyberbullies on Social Networks
Wu et al proposed a method for ranking nodes to identify trolls without using a PageRank algorithm.
Kumar proposed an iterative algorithm involving new decluttering operations and various centrality measures to detect trolls.
According to the Cyberbullying Research Center (CRC) [7], 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, cyberbullying contains hurtful messages which are present online for a long time. These messages can be accessed worldwide, and are often irrevocable. Laws about cyberbullying and how it is handled differ from one place to another.

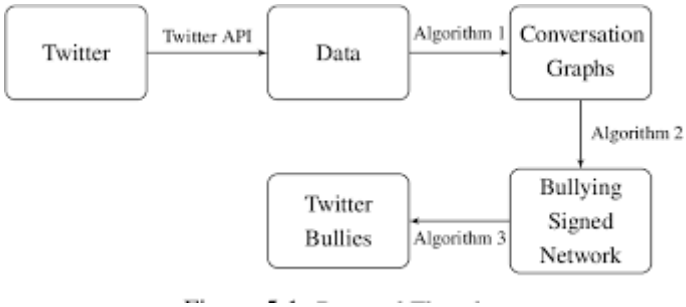


fig 2.1.1 Protocol Flowchart of BullyNet.

Social media platforms provide users with the option to self-report abusive behavior and content in addition to providing tools to deal with bullying. For example, Twitter has features that include locking accounts for a brief period of time or banning the accounts when the behavior becomes unacceptable. The body of work produced by the research community with regards to cyberbullying in social networks also needs to be expanded to get better insights and help develop effective tools and techniques to tackle the issue.

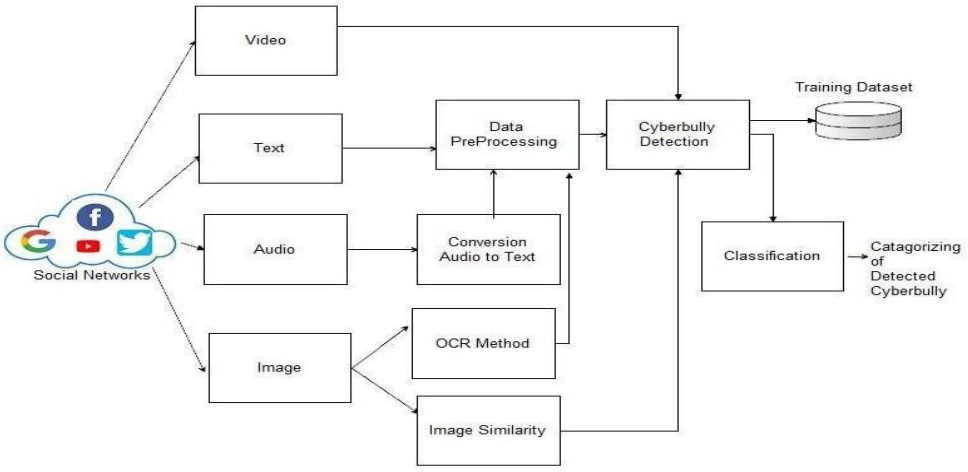


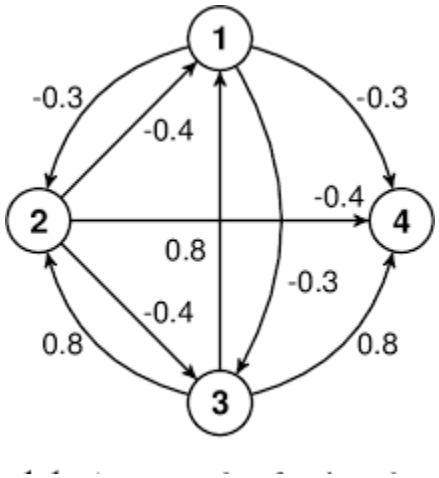
fig 2.1.2 Proposed Method for Cyberbully Detection Architecture

To identify cyberbullies in social media, we first need to understand how social media can be modeled. The common way of modeling relationship in social psychology [9] is to represent it as a signed graph with positive edge corresponds to the good intent and negative edge corresponds to malicious intent between people. Using the signed graph, we model the Twitter social network as a signed network to represent users' behavior [10] where nodes correspond to users, directed edges correspond to communications and/or relations between the users with assigned weight in the range [-1, 1], as illustrated in Figure.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703075 |



The large size and dynamic and complex structure of social media networks makes it challenging to identify cyberbullies. For example, on Twitter, hundreds of millions of tweets are sent every day on the social network platform.In this case we construct the social network as a graph and assign value based on the maliciousness of the user. Because the network analysis reduce the complex relationship between the users to the simple existence of nodes and edges.

III. METHODOLOGY

we first present an overview of the proposed three-phase bully finding algorithm and elaborate the steps in each phase. The objective of our solution is to identify the bullies from raw Twitter data based on the context as well as the contents in which the tweets exist. Given a set of tweets T containing Twitter features such as user ID, reply ID etc.
The proposed approach consists of three algorithms:
(i) Conversation Graph Generation Algorithm,
(ii) Bullying Signed Network Generation Algorithm,
(iii) Bully Finding Algorithm.
The first algorithm constructs a directed weighted conversation graph Gc 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 B to analyze the behaviour of users in social media.
The third algorithm consists of our proposed attitude and merit centrality measures to identify bullies from BComputer vision: Face detection, hand tracking, pose estimation, object tracking, scene segmentation, and more.

- 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.
- Then from the conversation graphs, a bullying signed network is generated using algorithm
- Finally, the bullies from Twitter are identified by applying algorithm.

3.1. Conversation Graph Generation:

The conversation graph generation algorithm 1, is constructed from a set of tweets T to generate directed weighted conversation graphs Gc for each conversation. The weights between the nodes or users are determined by analyzing the sentiment behind the text of a tweet and examining for curse words. We then provide a score based on the expression the text represents. For each tweet ti in T, the conversations are built by doing a binary search DID(ti) with the SID of the remaining tweets. If a match is found as t`, then it is appended with ti to form a new conversation. If a binary search match is found with an already existing tweet in a conversation ci then, ti is appended to tweets in ci .
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. There are different libraries or tools available to determine the sentiment of the content which includes sarcasm, emoji, images etc. Some of the them are: VADER, TextBlob, Python NLTK etc.
We use VADER (Valence Aware Dictionary and sentiment Reasoner), which is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It performs well with emojis,

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703075 |

emoticons, slangs and acronyms in a sentence. Cosine Similarity (CS) measures the similarity between two vectors using their inner product.

3.2. 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 2 is designed to determine the final outgoing edge weight, wij for the users in the conversation graphs Gc.

for every conversation graph gci, a bullying score S is calculated based on how a node/user interact with other nodes/users in the graph based on the tweet order (sorted in ascending order) i.e., tweets are arranged based on the conversation. For an edge e = (u, v), the bullying score Suv ≡ Iuv if the edge towards v is not a reply from u. Otherwise, the bullying score Suv is calculated as Iuv + α(Iuv − Svu) where α is a constant determined by the experiment as 0.6. Here, α is used to calculate how much percent of the difference between the sender and receiver should be taken to determine the bullying score S.

The bullying signed network graph B is constructed by merging all the conversation graphs Gc. 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 wuv. Step 4 outputs the bullying signed network graph B.

3.3. Bully Finding

This work, is to identify bullies from B using centrality measures. Since this paper 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 [34] a state of art method, that handles signed network 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 the 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 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.

Since merit is about whether the node is considered good or bad, it is calculated to be 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 truly it rates other nodes, it is calculated by the difference in the edge weight and the real trust of a node.

3.4. Data Sources

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, metadata (e.g., creation time, source ID, destination ID, reply/retweet, etc.) as well as information about the poster (e.g., username, followers, friends).

3.5. Data Volume and Availability

To prevent our own bias, we first randomly chose 5000 interconnected users and collected all the tweets in JSON format totaling 5.6M within a six IEEE Transactions on Computational Social Systems,Volume:8,Issue:2,Issue Date:April.2021 9 month time-frame between May and October 2017. We then extracted features like username, text, replyname, mentions and network based features like source ID, destination ID from the Twitter JSON. There were about 2% of the tweets which were in languages other than English.

3.6. Accuracy and Generalization

That is, with 5.7 million tweets dataset we did experiment for the tweets ranging from 1M, 2M, 3M, 4M and 5M for different α, β and γ values. 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 [41].

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703075 |

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

5. Parameter Formulas:

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

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

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

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

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

IV. RESULTS

Evaluation

This paper presents a novel framework of BullyNet 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 could 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 around 80% accuracy with 81% precision in identifying bullies for various cases.

	Precision	Recall	F1 Score
Chatzakou	75	53	79
Zhao	76.8	79.4	78
Singh	82	53	64
BullyNet	**77.6**	**77.6**	**79.4**

V. CONCLUSION

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. This paper presents a novel framework of Bully Net 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 could 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 around 80% accuracy with 81% precision in identifying bullies for various cases.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 3, March 2024 |

| DOI:10.15680/IJMRSET.2024.0703075 |

REFERENCES

1. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, “Detection of cyberbullying incidents on the instagram social network,” In Proceedings of the CoRR, 2015.
2. J.-M. Xu, X. Zhu, and A. Bellmore, “Fast learning for sentiment analysis on bullying,” in Proceedings of the First International WISDOM, 2012.
3. P. Galan-Garcia, J. De La Puerta, C. G´omez, I. Santos, and P. Bringas, “Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying,” 2014.
4 D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, “Mean birds: Detecting aggression and bullying on twitter,” in Proceedings of the ACM on WebSci, 2017.
5. L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, “Xbully: Cyberbullying detection within a multi-modal context,” in Proceedings of the Twelfth ACM International Conference on DataMining, 2019.

Impact Factor
7.521

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |