IDENTIFICATION AND CHARACTERIZATION OF CYBERBULLYING DYNAMICS IN AN ONLINE SOCIAL NETWORK

Dr. Subba Reddy Borra
Professor & Head of the
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous) Maisammaguda, Hyd-500100, Telangana, India.

Anusha Tenneti
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda, Hyd-500100, Telangana, India.

Neela Shanvitha
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous) Maisammaguda, Hyd-500100, Telangana, India.

Ogirala Rajeswari
Student
Department of Information Technology
Malla Reddy Engineering College for Women
(UGC-Autonomous)
Maisammaguda, Hyd-500100, Telangana, India.

Abstract*—This research applies a social network perspective to the issue of cyber aggression, or cyberbullying, on the social media platform Twitter. Cyber aggression is particularly problematic because of its potential for anonymity, and the ease with which so many others can join the harassment of victims. Utilizing a comparative case study methodology, the authors examined thousands of Tweets to explore the use of denigrating slurs and insults contained in public tweets that target an individual's gender, race, or sexual orientation. Findings indicate cyber aggression on Twitter to be extensive and often extremely offensive, with the potential for serious, deleterious consequences for its victims. The study examined a sample of 84 aggressive networks on Twitter and visualize several social networks of communication patterns that emanate from an initial, aggressive tweet. The authors identify six social roles that users can assume in the network, noting differences in these roles by demographic category. Serious ethical concerns pertain to this technological, social problem.

INTRODUCTION

Cyberbullying is a trendy topic nowadays because in this modern generation all the youth and students are fully addicted to social media like Instagram+, Facebook, Twitter, Telegram, etc., these are the common platforms where students and youth spend a lot of time when they are traveling, in their free time and someone makes social media an online business to sell their products and some other persons they like to discover, initiate, and get a fan following, etc., These are the most common things in our social media. During this corona time lot of people were addicted gaming's like battlefield - PUBG, and Free Fire, which the youth and children were mainly addicted to this. Some hackers where take this as an opportunity to harass them by using this platform as an opportunity to get the details by doing call records, photos, chats, and video calls. So, they harass them to get money and to get popularity.

Examples like:

Flaming: Heated online arguments and fights using vulgar and abusive language.
Harassment: Repeatedly sending cruel, offensive, or threatening messages.
Denigration: Exposing secrets of a person or gossip to damage the reputation of a person.
Impersonation: Breaking into the victim's account and sending emails.
Trickery: Tricking the victim into revealing sensitive information and passing it on to others.
Interactive Gaming: Most gaming consoles allow people to connect and play online providing a chance to abuse using chats and comments.

Due to the lack of existing datasets, a very few studies have been done on the detection of Cyberbullying. At present, whatever work has been done to prevent cyberbullying is not accurate and reliable. In this paper, we are going to review cyberbullying and the work. That has been done to detect cyberbullying. An article from The Times of India entitled "$188,776 Facebook grant for cyberbullying expert Sameer Hinduja". Clearly states the increasing cyberbullying from the fact that Sameer Hinduja, an Indian-American and cyberbullying expert from Florida Atlantic University, has received a $ 188,776 grant from the social networking site Facebook to study cyberbullying. The goal of the study is to study the nationwide prevalence and scope of cyberbullying. According to Ipsos –a global market research company has found that 3 out of 10 parents in India say that their children have been victims of cyberbullying, majorly through Facebook and Orkut. The frequency of cyberbullying in India was found to be very high with 32% of children having access to the Internet or mobile phones. An article from. The Indian Express Alarming! 50% of Indian youths who have experienced cyberbullying found that most Indian parents don't find it important to talk to their children about online safety. Although there is an age restriction on joining various social networking sites 10-12 years teens report very high access to these sites. The Global Youth Online Behaviour Survey ranked India third in cyberbullying.

Nishant (name changed), outshone his school seniors in basketball. But suddenly, he did not want to play and became withdrawn from the game. He finally confided in his father that his school seniors had been harassing him online since they could not pull him down on the court. So his father went to a counsellor. National Crime Prevention Council defines cyberbullying as sending text or images to hurt or embarrass another person by using the Internet, mobile phones, or other devices. According to research conducted by Symantec, only 25% of the parents were aware that their child was involved in cyberbullying incidents. According to a survey majority of cyberbullying is done through Facebook and around 55% of the youth exposed to cyberbullying committed suicide. Cyberbullying incidents don't come as a surprise. Schoolchildren confirm that cyberbullying is becoming the easiest way to get back at someone. A person can be knocked down in front of a large number of people online. Many cyberbullies think that bullying others online is funny. Cyberbullies may not realize the consequences for themselves of cyberbullying. The things teens post online now may reflect badly on them later in the future. Also, cyberbullies and their parents may face legal charges for cyberbullying. Teens may think that if they use fake**.**

Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. Cyberbullying is unique in the degree to which it provides anonymity and in its ability to Facilitate the participation of multiple individuals in the harassment of victims. Perhaps for these reasons, victims often exhibit emotional distress (Ybarra, Mitchell, Wolak, & Finkelhor, 2006), Low self-esteem (Patchin & Hinduja, 2010), loneliness (Sahin, 2012), and other negative emotions (e.g., Juvonen & Gross, 2008). Those targeted by forms of electronic aggression also reported more suicidal thoughts and were more likely to attempt suicide than those who had not been victimized (Hinduja & Patchin, 2010). The purpose of this research is to examine cyber aggression on the social media website of Twitter. On this widely popular venue, Twitter enables users to send and read short, informative messages called "tweets" on a website with millions of active users each day. Yet Twitter can be used by some to disseminate aggressive, bullying messages, and the website has come under scrutiny for some of the most public instances. With only a few exceptions, bullying on Twitter has received little attention in scholarly literature. In this project, therefore, the authors study instances of aggression. Tweets that derogate individuals on the basis of one or more of three demographic characteristics: gender, race, and/or sexual orientation. The authors examine the network spread of cyber aggression within a Twitter conversation and identify the social roles of the participants within the interchange. Given the potentially serious, ethical questions raised by cyberbullying (Neves & Peinhero, 2010), studying this type of damaging interchange on the relatively novel technological site of Twitter remains particularly important. Predictions made and feature analysis. Finally, AugmentED is examined using a real-world dataset of 156 college students.

EXISTING SYSTEM

Hsien used an approach using keyword matching, opinion mining, and social network analysis and got a precision of 0.79 and recall of 0.71 from datasets from four websites. Patxi Gal´an-Garc´1a et al proposed a hypothesis that a troll(one who cyberbullies) on social networking sites under a fake profile always has a real profile to check how others see the fake profile. They proposed a Machine learning approach to determine such profiles. The identification process studied some profiles which have some kind of close relation to them. The method used was to select profiles for study, acquire information on tweets, select features to be used from profiles, and use ML to find the author of tweets. 1900 tweets were used belonging to 19 different profiles. It had an accuracy of 68% for identifying the author. Later it was used in a Case Study in a school in Spain where out of some suspected students for Cyberbullying the real owner of a profile had to be found and the method worked in the case. The following method still has some shortcomings. For example, a case where trolling account doesn't have a real account to fool such systems or experts who can change writing styles and behaviors so that no patterns are found. For changing writing styles more efficient algorithms will be needed. Mangaonkar et al. proposed a collaborative detection method where there are multiple detection nodes connected to each other where each node uses either a different or the same algorithm and data and results were combined to produce results. P. Zhou et al.[4] suggested a B-LSTM technique based on concentration. Banerjee et al.[5]. used KNN with new embeddings to get a precision of 93%. Where there are multiple detection nodes connected to each other where each node uses either a different or the same algorithm and data and results were combined to produce results. P. Zhou et al. suggested a B-LSTM technique based on concentration. Banerjee et al. used KNN with new embedding to get a precision of 93%.

Disadvantages –

- A vocabulary is not designed from all the documents. The vocabulary may consist of all words (tokens)
- In all documents or some top frequency tokens, the øTf-If method is not similar to the bag of words model since it uses the same way to create.
- A vocabulary to get its features.

PROPOSED SYSTEM

In the proposed system Cyberbullying detection is solved in this project as a binary classification problem where we are detecting two major's form of Cyberbullying: hate speech on Twitter and Personal attacks on Wikipedia and classifying them as containing Cyberbullying or not.

Tokenization: In tokenization, we split the raw text into meaningful words or tokens. For example, the text less than we will do it equals to can be tokenized into 8we9, 8will9, 8do9, 8it9. Tokenization can be done into words called word tokenization or sentences called sentence tokenization. Tokenization has many more variants but in the project, we use Regex Tokenizer. In regex, tokenizer tokens are decided based on a rule which in this case is a regular expression. Tokens matching the following regular expression are chosen For Example for the regular expression 8\w+9 all the

alphanumeric tokens are extracted. Stemming: Stemming is the process of converting a word into a root word or stem. For example, for three words 8eating9 8eats9 8eaten9, the stem is 8eat9. Since all three branch words of root 8eat9 represent the same thing they should be recognized as similar. NLTK offers 4 types of stemmers: Porter Stemmer, Lancaster Stemmer, Snowball Stemmer, and Reg exp Stemmer. The following project uses Porter Stemmer Stop word Removal: Stop words are words that do not add any meaning to a sentence example Some stop words for the English language are: what is, at, a, etc. These words are irrelevant and can be removed. NLTK contains a list of English stop words that can be used to filter out all the tweets. Stop words are often removed from the text data when we train Deep learning and Machine learning models since the information they provide is irrelevant to the model and helps in improving performance.

Advantages –

The common Bag of Words model takes as input multiple words and predicts the word based on the context.

- Input can be one word or multiple words. CBOW model takes a means of the context of input words but
- Two semantics can be clicked for a single word. i.e. two vectors of Apple can be predicted. First is for the firm Apple and next is Apple as a fruit.

SYSTEM ENVIRONMENT

Python:
It is an interpreter, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python interpreters are available for many operating systems. Python is managed by the non-profit Python Software Foundation. Python features a dynamic types system and automatic memory management. It supports multiple programming paradigms, including object-oriented, functional, and procedural, and has a large and comprehensive standard library. Python is an easy-to-learn yet powerful and versatile scripting language, which makes it attractive for Application Development.

Python IDLE:
IDLE stands for Integrated Development and Learning Environment. The story behind the name IDLE is similar to Python. Guido Van Rossum named Python after the British comedy group Monty Python while the name IDLE was chosen to pay tribute to Eric Idle, who was one of Monty Python's founding members. IDLE comes bundled with the default implementation of the Python language since the 01.5.2b1 release. It is packaged as an optional part of the Python packaging with many Linux, Windows, and Mac distributions.

IMPLEMENTATION

A. Training and Testing Dataset

In this module, we will focus on the dataset that we have gathered, eliminating all rows with null entries at first. At that point, we will dispose of any unnecessary features that could imperil our algorithm's accuracy. Here we will also divide the dataset into two sections - Training and Testing. 80% of the dataset will be utilized for training the model and the rest 20% will use for checking the training model's precision. The data gathered is manually labelled as either bully (sexual, threat, troll, or religious) or not-bully. Along with it, the dataset also has three other columns specifying the category of the comments passed, the gender on which the comment is made, and the total number of reactions for each comment.

B. DATA PRE-PROCESSING

The data collected had to be pre-processed since it had traces of unstructured content. It meant we needed to clean or trim the data to obtain higher accuracy. Various steps needed to be followed for pre-processing the data such as data cleaning, stop word removal, and tokenization. With the help of a stop word filter, we deleted any needless words in all the text conversations in line with the Bengali vocabulary. The term stop words mean those words that don't give any helpful data to decide in which category a text should be classified. For facilitating the further processes with the motive of not distinguishing between capital letters and lowercase letters, we transformed the whole data into lowercase. Furthermore, tokenization had to be practiced on these text contents to facilitate the feature extraction step. Tokenization can be defined as a way of separating or isolating every word that compiles in a document or even a conversation.

C. FEATURE EXTRACTION

The pre-processed data with text conversations will be converted into a vector space model where these text conversations will be described with a vector of extracted features using the Term Frequency Inverse Document Frequency (TFIDF). TFIDF is used for measuring or evaluating how relevant a word is to a document or a collection of documents. Thus, the main aspect of TFIDF is that it performs well on the text and gets the weights of these words regarding the document or the sentence. Along with TFIDF, we will also use word-level feature extraction methods; this specific strategy is known as Bag of Words or < Bag of n-grams= representation. It implies that documents are defined or represented by occurrences of the words while completely neglecting the position or order of the words in the document. There is various parameter that is mostly used to combine the vectorizer and a machine learning model, one such parameter is max df used to remove terms that appear too frequently in the document.

CLASSIFICATION

The final step in the proposed model is classification, where the features extracted are put into an algorithm to train and test the classifier and hence to determine whether it can successfully detect cyberbullying or not.

We will use various machine learning methods, and algorithms such as Support Vector Machine (SVM), Logistic Regression
(LR), Random Forest and Passive Aggressive (PR) classifier. The assessment of all these classifiers is completed utilizing a few assessment lattices. Among those criteria are Accuracy, precision, recall, and f-score.

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 a better understanding of the relationships between users in social media, and 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 the signed network, we achieved around 80% accuracy with 81% precision in identifying bullies for various cases. There are still several open questions deserving 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 a 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 proximity increase bullying behavior?

REFERENCES

[1] J. Tang, C. Aggarwal, and H. Liu, “Recommendations in signed social networks,” in Proceedings of the International Conference on WWW, 2016, pp. 31–40.

[2] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem forsocial networks,” Proceedings of the ASIS&T, vol. 58, no. 7, pp. 1019–1031, 2007.

[3] U. Brandes and D. Wagner, “Analysis and visualization of social networks,” in Graph drawing software, 2004, pp. 321–340.

[4] X. Hu, J. Tang, H. Gao, and H. Liu, “Social spammer detection with sentiment information,” In Proceedings of IEEE ICDM, pp. 180—189, 2014.

[5] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, Trolls just want tohave fun, 2014, pp. 67:97–102.

[6] S. Kumar, F. Spezzano, and V. Subrahmanian, “Accurately detectingtrolls in slashdot zoo via decluttering,” in Proceedings of IEEE/ACM ASONAM, 2014, pp. 188–195.

[7] J. W. Patchin and S. Hinduja, “2016 cyberbullying data,” 2017.

[8] C. R. Center, https://cyberbullying.org/bullying-laws.

[9] D. Cartwright and F. Harary, “Structural balance: a generalization of heider’s theory.” Psychological review, vol. 63, no. 5, p. 277, 1956.

[10] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Signed networks insocial media,” in Proceedings of the SIGCHI CHI, 2010, pp. 1361–1370.

[11] R. Plutchik, “A general psychoevolutionary theory of emotion,” in Theories of emotion, 1980, pp. 3–33.

[12] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithmsand applications: A survey,” Proceedings of the Ain Shams engineering journal, vol. 5, no. 4, pp. 1093–1113, 2014.

[13] L. Tang and H. Liu, “Community detection and mining in social media,” Synthesis lectures on data mining and knowledge discovery, vol. 2, no. 1, pp. 1–137, 2010.

[14] S. Bhagat, G. Cormode, and S. Muthukrishnan, “Node classification insocial networks,” in Social network data analytics, 2011, pp. 115–148.

[15] J. Tang, Y. Chang, C. Aggarwal, and H. Liu, “A survey of signed networkmining in social media,” In Proceedings of the ACM Comput. Surv., no. 3, pp. 42:1–42:37, 2016.

[16] J. Kunegis, J. Preusse, and F. Schwagereit, “What is the added valueof Cochrane Database of Systematic Reviews, vol. 4, no. 4, pp. CD008958, 2014. negative links in online social networks?” in Proceedings of the International Conference on WWW, 2013, pp. 727–736.

[17] Z. Wu, C. C. Aggarwal, and J. Sun, “The troll-trust model for ranking insigned networks,” in Proceedings of the ACM International Conference on WSDM, 2016, pp. 447–456.

[18] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” in Proceedings of the ICDCN, 2016.

[19] V. K. Singh, Q. Huang, and P. K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," In Proceedings of the IEEE/ACM ASONAM, pp. 884—-887, 2016.

[20] 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.

[21] J.-M. Xu, X. Zhu, and A. Bellmore, "Fast learning for sentiment analysison bullying," in Proceedings of the First International WISDOM, 2012, pp. 10:1–10:6.

[22] A. C. Squicciarini, S. M. Rajtmajer, Y. Liu, and C. H. Griffin, "Identificationand characterization of cyberbullying dynamics in an online Social network," in Proceedings of the IEEE/ACM ASONAM, 2015, pp.280–285.

[23] 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," vol. 24, pp. 42–53, 2014.

[24] 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, pp. 13–22.

[25] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in Proceedingsof the Twelfth ACM International Conference on Web Search andDataMining, 2019, pp. 339–347.