ISSN: 2366-1313

CYBER BULLYING RECOGNITION ON SOCIAL MEDIA USING MACHINE LEARNING

[1]P.V MADHAVI, [2]K.RAJ KUMAR

[1,2]Assistant Professor, Dept of CSE, Megha institute of engineering and Technology for women, Ghatkesar (T.S)

Abstract: *Cyberbullying is one of the main problems found on the Internet that affects teenagers and adults alike. This has led to false incidents like suicide and depression. Content regulation on social media platforms has become a growing necessity. The following study uses data from specific types of cyberbullying, hate tweets from Twitter, and comments based on personal attacks from Wikipedia forums to analyze cyberbullying in text information using natural language processing and machine learning. An index-based version can be created. Three strategies for feature extraction and four classifiers are studied to illustrate the high-quality approach. For tweet records, the version provides over ninety percent accuracy and for Wikipedia information it provides over 80% accuracy.*

Keywords: *Cyberbullying, social media, BERT, NLP, Machine learning, Twitter*

I. INTRODUCTION

Now more than ever, technology has become an indispensable part of our lives. As the network evolves. Social network is a trend nowadays. But like all other cases, abusers sometimes come out late, but something positive can happen. Now, cyberbullying is not uncommon these days. Social networking sites are great tools for communication between people. The use of social media has become a lot over the years, although people usually find unethical and unethical ways to do horrible things. We see this happening in teenagers or occasionally young adults. One of their vices is bullying others online. In an online environment we cannot easily tell if someone is saying something just for a laugh or if their intentions may be different. Often, there is only the possibility of laughing at a funny, "or don't take it so seriously" story. Cyberbullying is the use of race to annoy, threaten, embarrass or attack another man or woman. Often, this online fighting results in real threats to some people's lives. Some people have resorted to suicide. It is important to anticipate these types of games from the beginning. Measures can be taken to prevent this, for example, if a person's tweet/post is deemed offensive, their account can be terminated or suspended for a certain period. On Instagram, when someone shares photos and videos made by the user to be uncomfortable, so the user can monitor or block them. Users can also report a violation of our community and make Recommendations to the app. While these platforms provide an opportunity for people to interact and communicate in ways that were previously unimaginable, they have also given rise to negative behaviours like cyberbullying. Cyberbullying is the act of intimidating, threatening, or coercing others through the internet using digital or electronic means such as social media, email, text messaging, blog postings. Cyberbullying, also known as internet harassment, frequently makes use of insulting, hostile, or threatening language. Cyberbullies frequently hide their true identities behind fake digital profiles [2].

Cyberbullying is a major and widespread problem in today's digital culture that affects a growing number of Internet users, particularly impressionable teenagers and young people. In a way, unlike its digital equivalent, which can happen anytime, anywhere with only a few keystrokes on a keyboard, physical bullying is relatively restricted to specific locations or periods of the day.

Cyberbullying is a form of psychological abuse that has a big influence on society. Events of cyberbullying have been rising, especially among young individuals who spend the majority of their time switching between various social media sites. Because of their popularity and the anonymity that the Internet offers to abusers, social media networks like Twitter and Instagram are particularly vulnerable. Cyberbullying may even result in severe mental disorders and detrimental impacts on mental health. The majority of suicides are caused by the worry, depression, stress, and social and emotional challenges brought on by instances of cyberbullying [3].

These issues lead to the creation of techniques and tools for the early identification and prevention of such abusive behaviour, particularly when it develops on social media platforms. Developing efficient and effective strategies for detecting such online occurrences involves many complexities. This highlights the need for a method to spot cyberbullying in messages posted on social media (e.g., posts, tweets, and comments). The key tasks in addressing cyberbullying risks are the detection of cyberbullying events from tweets and the implementation of preventive measures. This is because cyberbullying is increasingly an issue on Instagram. Therefore, there is a larger need to conduct more study on social network-based CB in order to gain more knowledge and contribute to the creation of tools and strategies that will successfully tackle the problem.

The main methods for detecting cyberbullying on the Instagram platform are comment categorization and, to a lesser extent, topic modelling techniques. Text categorization using supervised machine learning (ML) models is frequently used to separate bullying-related and non-bullying comments. Bullying and non-bullying tweet classification has also been accomplished using deep learning (DL) based classifiers. Only a predetermined set of events may be adequate for supervised classifiers; however, they are unable to handle dynamically changing comments. The method of extracting the crucial subjects from a piece of data to create the patterns or classes in the entire dataset has long been topic modelling methodologies. Despite the similarity in principle, short texts cannot be effectively covered by standard unsupervised topic models; as a result, specialized unsupervised short text topic models were used. These models successfully extract the trending topics from comments and hashtags for additional processing. By utilising the bidirectional processing, these models aid in the extraction of significant issues. However, in order to get sufficient prior information for these unsupervised models, significant training is required, which is not always sufficient. Given these restrictions, a successful strategy for classifying comments and hashtags must be created in order to fill the gap between the classifier and the topic model and greatly improve flexibility.

II. LITERATURE SURVEY

Lot of research have been done to find possible solutions to detect Cyberbullying on social networking sites.

In [4] Hsien et al 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´ıa et al.[5] proposed a hypothesis that a troll(one who cyberbullies) on a social networking sites under a fake profile always has a real profile to check how other see the fake profile. They proposed a Machine learning approach to determine such profiles. The identification process studied some profiles which has some kind of close relation to them. The method used was to select profiles for study, acquire information of tweets, select features to be used from profiles and using ML to find the author of tweets. 1900 tweets were used belonging to 19 different profiles. It had an accuracy of 68% for identifying 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 doesnt have a real account to fool such systems or experts who can change writing styles and behaviours so that no patterns are found . For changing writing styles more efficient algorithms will be needed. Mangaonkar et al. [6] proposed a collaborative detection method where there are multiple detection nodes connected to each other where each nodes uses either different or same algorithm and data and results were combined to produce results. P. Zhou et al.[7] suggested a B-LSTM technique based on concentration. Banerjee et al. used KNN with new embeddings to get an precision of 93%. Kelly Reynolds, April Kontostathis and Lynne Edwards[8] propose a Formpring(A forum for anonymous questions-answers) dataset which gives recall of 78.5% using Machine learning Algorithms and oversampling due to imbalance in cyberbullying posts Jaideep Yadav, Kumar and Chauhan used a latest language model developed by google called BERST which generates contextual embeddings for classification. The model gave a F1 score of 0.94 on form spring data and 0.81 on Wikipedia data.Maral Dadvar and Kai Eckert trained deep neural networks on Twitter,Wikipedia and Formspring datasets and used the model on Youtube dataset for the same and achieved F1 score of 0.97 using Bidirectional Long Short-Term Memory(BLSTM) model.Sweta Agrawal and Amit Awekar [9] used similar same datasets for training Deep Neural Networks but one of its key focus is swear words and their use as features for the task. They determined how the vocabulary for such models varies across various Social Media Platforms.Yasin N. Silva,Christopher Rich and Deborah Hall [10] built BullyBlocker,a mobile application that informs parents of cyberbullying activities against their child on Facebook which counted warning signs and vulnerability factors to calculate a value to measure probability of being bullied.

In [11] we present Mal JPEG, a machine learning-based solution for efficient detection of unknown malicious JPEG images. To the best of our knowledge, we are the first to present a machine learning-based solution tailored specifically for the detection of malicious JPEG images. Mal JPEG features are extracted based on the structure of the JPEG image. Mal JPEG features were defined based on an understanding of how attackers use JPEG images in order to launch attacks and how it affects the JPEG file structure in comparison to regular benign JPEG images. The features are simple and relatively easy to extract statically (without actually viewing the image) when parsing the JPEG image file. This paper presents a robust methodology to distinguish bullies and aggressors from normal Twitter users by considering text, user, and network-based attributes. Using various state-of-the-art machine learning algorithms, these accounts are classified with over 90% accuracy and AUC. Finally, the current status of Twitter user accounts marked as abusive by our methodology, and study the performance of potential mechanisms that can be used by Twitter to suspend users in the future. The drawback of this paper is the average level performance provided by the state-of-the-art machine learning algorithm and it is susceptible to errors. The paper did not provide real-time detection of abusive behaviors with the aid of properly tuned distributed stream and parallel processing engines. It did not repeat the same analysis on other online social media platforms such as Facebook, Foursquare, and YouTube, in order to understand if the provided methods can detect similar behavioural patterns and can help bootstrap their effort to combat them.

Amanpreet Singh et al. [9] has reviewed many previous research papers related to machine learning models, pre-processing techniques, evaluation of machine learning models, etc. This paper includes study research based on various previous research papers. They've discussed used methodology, datasets, conclusions/findings, content-based features, demerits, technique and used models, pre-processing steps used for the model. For, researching purposes, they've explored Scopus and the IEEE Xplore virtual library, ACM Digital Library. Using citations, 51 academic papers were discovered. Based on concluding arguments, abstracts, and titles, 18 papers were found not to apply to the survey so 18 papers were discarded. In this paper for the survey, they've reviewed 27 papers from 33 papers after filtration. In, each of the 27 research papers binary classification is used for cyberbullying detection. And most of them have used the Support Vector Machine (SVM) algorithm for detection.

III. PROPOSED SYSTEM

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

A. Twitter Dataset

The Twitter Dataset is combined from two datasets containing hate speech :

- Hate Speech Twitter Dataset by Waseem, Zeerak and Hovy, Dirk[11] which contains 17000 tweets labelled for sexism or racism. The tweets are mined using the annotations .5900 tweets are lost due to accounts being deactivated or tweet deleted.

- Hate Speech Language Dataset by Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar.

It contained 25000 tweets obtained by crowdsourcing. This gives total 35787 tweets for the task distribution for which is shown in Fig. 3.For the following dataset, 70 percent(25,050) of this dataset is used as training data and 30 percent as testing data(10,737) .

B. Wikipedia Dataset

The Wikipedia dataset by Wulczyn, Thain and Dixon[13] contains 1M comments labelled for Personal attacks.For the analysis 40000 comments are used from the dataset from which 13000 comments are labelled as Cyberbullying due to personal attack. These comments are extracted from conversations between editors of pages on Wikipedia labelled by 10 annotators via Crowd Flower. For this dataset the same split(70 percent i.e 28000 to training data and 30 percent i.e 12000 to testing data) is used.Fig. 2 shows its distribution.

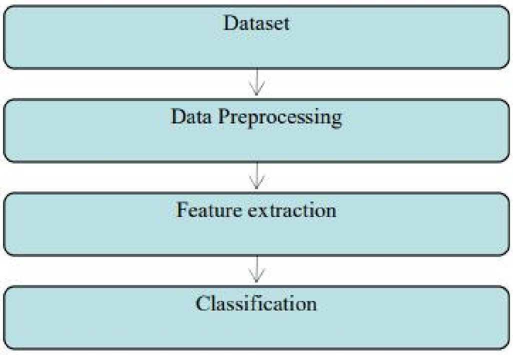

Fig.1 Proposed method architecture

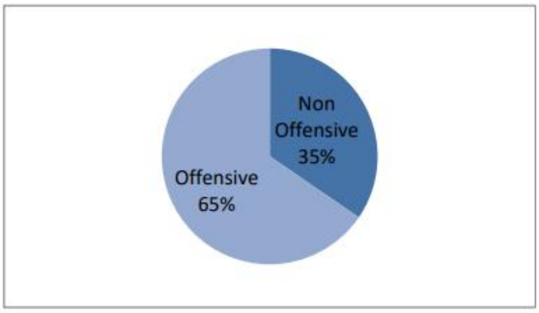

Fig.2 Distribution of Tweets in Twitter Dataset.

Bag of Words model

The BoW that is bag of words model is a simple method of extracting features from documents that uses occurrence of words within a document. Bag of Words model has two important parts: • A vocabulary of words(tokens) derived from all documents

• A way of measuring all these words as features in each document It is referred to as ‘bag’ because the model only concerns with the word rather than its order of occurrence in the document. The intuition for this method is that similar documents have similar words in them. The Bag of Words model uses the following procedure:A vocabulary is designed from all the documents. The vocabulary may consist of all words (tokens) in all documents or some top frequency tokens e.g. top 10 features with max occurrences in the corpus. Also features can be extracted for vocabulary in multiple forms based on number of words used per feature. e.g. for the sentence ‘This was the best ever’.

IV. CONCLUSION

Cyber bullying across internet is dangerous and leads to mishappenings like suicides, depression etc and therefore there is a need to control its spread. Therefore cyber bullying detection is vital on social media platforms. With avaibility of more data and better classified user information for various other forms of cyber attacks Cyberbullying detection can be used on social media websites to ban users trying to take part in such activity In this paper we proposed an architecture for detection of cyber bullying to combat the situation.We discussed the architecture for two types of data: Hate speech Data on Twitter and Personal attacks on Wikipedia.For Hate speech Natural Language Processing techniques proved effective with accuracies of over 90 percent using basic Machine learning algorithms because tweets containing Hate speech consisted of profanity which made it easily detectable. Due to this it gives better results with BoW and Tf-Idf models rather than Word2Vec models However, Personal attacks were difficult to detect through the same model because the comments generally did not use any common sentiment that could be learned however the three feature selection methods performed similarly.Word2Vec models that use context of features proved effective in both datasets giving similar results in comparatively less features when combined with Multi Layered Perceptrons.

REFERENCES

1. F. Mishna, M. Khoury-Kassabri, T. Gadalla, and J. Daciuk, “Risk factors for involvement in cyber bullying: Victims, bullies and bully–victims”
2. K. Miller, “Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law’s limited available redress”
3. A. M. Vivolo-Kantor, B. N. Martell, K. M. Holland, and R. Westby, “A systematic review and content analysis of bullying and cyber-bullying measurement strategies”
4. H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, “Associations between

cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren"

5. M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context".
6. G. A. León-Paredes et al., Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language, CHILECON pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684. (2019)
7. P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, A Framework for Hate Speech Detection Using Deep Convolutional Neural Network, in IEEE Access, vol. 8, pp. 204951-204962,, doi: 10.1109/ACCESS.2020.3037073. (2020).
8. S. M. Kargutkar and V. Chitre, A Study of Cyberbullying Detection Using Machine Learning Techniques, ICCMC, pp. 734-739, doi: 10.1109/ICCMC48092.2020.ICCMC-00013 7. (2020)