ZKG INTERNATIONAL ISSN: 2366-1313

BULLYNET: REVEALING CYBERBULLIES ON SOCIAL MEDIA PLATFORMS

#1BONTHA SAI KIRAN RAO,
#2EREDDY ROHITH,
#3SHAGUFTHA BASHEER, *Assistant Professor,*
Department of Computer Science and Engineering,
SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, TS.

ABSTARCT: Cyberbullying occurs when someone is intimidated using technology while online. Criminals can readily discover targets on social networking platforms and then harass and abuse young adults who are unable to protect themselves. Using machine learning, we can identify threats' language patterns and devise methods to detect abusive material online. Most of the research on utilizing machine learning to detect cyberbullying has focused on languages such as English, Arabic, and Chinese. There isn't much writing about regional Indian languages. Our findings suggest that our system may detect cyberbullying literature published in an Indian language that is not widely spoken or in specific places.

Keywords: Cyberbullying, machine learning, random forest, passive aggressive classifiers.

1. INTRODUCTION

Cyberbullying is not the same as normal harassment, yet it is still extremely distressing. It has the same consequences and risks as before, if not worse, and will persist longer. Although cyberbullying occurs on the internet rather than in real life, it should still be considered wrong. Harassment can take several forms depending on the circumstances. Being a fraud or hacking into someone's page is not what it truly entails. It also includes making negative remarks about someone online or spreading stories about them in order to criticize them. Cyberbullying, often known as "social media bullying," occurs when someone manipulates, harasses, or mocks another person online. These heinous acts are extremely destructive and can have a serious and negative impact on anyone. They typically occur on the internet, in public locations, and on other websites.

PURPOSE

Cyberbullying has increased significantly during the previous ten years. This type of harassment is not limited to English; it can also occur in other languages. Cyberbullies want huge groups of people, so it's critical to understand the differences between cyberbullying in other languages. This means that the algorithm we're developing will aid in the discovery of cyberbullying content in Bengali, a widely spoken and regional language in India. There are few works in this language that address cyberbullying. We will utilize the same cyberbullying approaches as have already been applied in English language literature. As a result, semantic discrepancies between English and non-English material may cause variations in performance and execution. To address these issues, our concept proposes employing machine learning techniques and user data to detect digital abuse in text.

EXISTING SYSTEM

Hsien used an approach using keyword matching, opinion mining and social network analysis and got a precision Hsien employed a technique that incorporated phrase matching, opinion mining, and social network analysis. He achieved a precision of 0.79 and a recall of 0.71 using dataset

ISSN: 2366-1313

websites.Patxi from Galõan-Garcñıa et al. proposed that trolls (cyberbullies) on social media should have a real profile to understand how others perceive the phony profile. They devised a method to find these profiles using machine learning. During the identification procedure, certain names associated with them were examined.

The processes followed were to select profiles to examine, gather information on tweets, select attributes to use from profiles, and use machine learning to identify tweet writers. 1900 tweets from 19 distinct accounts were used. At 68% accuracy, it was able to identify the author. Following that, it was utilized in a Case Study at a Spanish school to determine the true owner of a student's profile who had been accused of cyberbullying. The method was effective in the circumstance. The following method is still not great in certain areas.

For example, a trolling account may not have an actual account in order to mislead these algorithms, or experts may alter their writing and behavior so that no trends are detected. We will need more efficient systems to change our writing styles.

Mangaonkarproposed a joint detection approach in which many detection nodes are linked together and employ either the same or alternative algorithms. The data and outcomes from each node are then merged to produce results. P. Zhou et al. suggested a B-LSTM concentration-based method. Banerjee et al. achieved an accuracy of 93% when they employed KNN with novel embeddings.

DISADVANTAGES

- A vocabulary does not consist of all papers. The vocabulary could include all of the words (tokens) used in all papers, or only the most common ones.
- The Tf-Idf technique is not the same as the bag of words model, despite the fact that it generates a vocabulary in the same way to identify its characteristics.

2. LITERATURE SURVEY

The researchers put together the dataset in the article using a Java app that was made to separate Bangla text messages from those sent on social media sites like Facebook and Twitter. Along with the conversations, they manually marked up the Twitter Rest API to get data on client segments, and they added a "bag of words" way to the model. After the model was made, they checked how well the descriptions of the different machine learning methods (like SVM, KNN, Naive Bayes, and J48) fit it. There were two parts to the evaluations. The first part used content discussions in Bangla Text to build and test a model.

The second part added user data and content-based highlights. SVM did better than a lot of other computations in both stages. Paper is one of the few papers that has tried to find harmful Bangla text. It also suggests using unigram string highlighting along with a root level calculation to find harmful information in order to get a better result. They tried out different string qualities, like bigram, trigram, and unigram, to find the best highlight for their suggested solution. The features of unigrams don't look at how important words are in a single text. Still, this part mostly draws attention to terms that were more damaging. The authors of the study put together a dataset of text conversions and comments from the comments sections of famous Facebook posts . There were two different groups that the dataset was split into: bullying and not bullying. Within the bully group, four subcategories were made clear: sexual, troll, religious, and threat. The dataset also had a gender classification appendix with details about the author and the person who was supposed to receive the data. The number of responses to each comment was added to the collection of comments, along with some extra information about the person, like what they do for a living. They made the information more useful by giving a report for each column in the dataset.

The large amount and variety of data found in each group makes it possible to build a strong machine

ISSN: 2366-1313

learning model that can effectively spot different types of cyberbullying in the Bangla language. Machine Learning techniques can be used to find instances of offensive authoritarianism in language and then build a model that can tell the difference between different kinds of online harassment. The main point of the study was to create a controlled machine learning model that could tell the difference between English-language online harassment and stop it. Digital harassment content from Kaggle was used as a sample to test the model. The effect of SVM and neural network classifiers on TFIDF and methods for extracting sentiment was thought about. The Neural Network classifier did better than the SVM classifier in terms of performance.

The Neural Network model also did better than their predictors in terms of accuracy and f-score when this study was compared to a similar one that used the same dataset. In their work, the authors suggested a model that can be used to find cyberbullying in more than one Indian language, mostly in Marathi and Hindi. For the information, they made their own API and scraper to get it from newspaper reviews, trip reviews, and social media sites. The dataset was put into the model after being labeled by hand and having stop words and special characters taken out. It used a number of machine learning languages, such as Multinomial Naive Bayes, Stochastic Gradient Descent, and Logistic Regression. The information was put into groups using the "bag of words" method, which counts how often words are used without looking at their grammar, order, or how often they appear. After figuring out the f1 scores for each ML algorithm, they found that LR has the lowest error rate of all of them, they gave an overview of how to recognize cyberbullying in more than one language. To make their model stand out, they tried to find cyberbullying in Arabic. This was because, according to their thorough research, most work on finding cyberbullying had been done in English. In their work, they used a number of machine learning methods to find cases of cyberbullying. They got 32,000 tweets for the dataset, and about 1800 were found to be threatening.

A method called Support Vector Machine (SVM) and another one called Naïve Bayes helped the researchers find cases of cyberbullying with 92% and 90% accuracy, respectively. The results were not perfect when this framework was compared to older models that were used to find English abuse. The point of this work was to show how easy it is to spot cyberbullying in Arabic. This makes it more likely that abuse can also be found in other regional or unusual languages.

3. PROPOSED SYSTEM

The two most common types of cyberbullying in this study are personal threats on Wikipedia and hate speech on Twitter. These are labeled as either cyberbullying or not. Cyberbullying tracking is thought of as a problem of two types.

Tokenization turns plain text into phrases that make sense. These phrases are called tokens. To give you an example, the phrase "we will do it" can be broken down into "it," "we," "will," and "do." There are two types of tokenization processes: word tokenization and sentence tokenization. For our project, we use Regex Tokenizer, but there are many other ways to tokenize data as well. Regex tokenizers use a rule, in this case a regular expression, to decide what tokens to use. Tokens that match the given regular phrase are chosen. Like When the regular phrase '\w+' is used, every alphanumeric token is taken out.

The act of stemming a word is to find its stem or root word. The word "eat" is at the root of all three words "eating," "eats," and "eaten." The three words that come from the root "consume" all mean the same thing, so they should all be thought of as synonyms. Porter Stemmer, Lancaster Stemmer, Snowball Stemmer, and Regexp Stemmer are the

ISSN: 2366-1313

four types of stemmers that NLTK has to offer. Porter and Stemmer are used in the next project.

ADVANTAGES

- Stop words are words that add nothing to the meaning of a sentence. "What," "is," "at," "a," and other words like these are stop words in English. These sentences aren't needed and can be taken out.
- You can use the English stop words that come with NLTK to get rid of all tweets. To make machine learning and deep learning models work better, stop words are often taken out of the training text. This is done because the information they carry is not thought to be useful.

4. IMPLEMENTATION

A. Training and Testing Dataset Training and Evaluation Set. This module will focus on the dataset we've obtained, first deleting any rows with null entries. At that point, we will delete any unnecessary features that may endanger the accuracy of our algorithm. In this case, the dataset will be divided into two sections: training and testing. We will train the model with 80% of the dataset and assess its accuracy with the remaining 20%. The obtained data is manually classed as non-bully or bully (sexual, threat, troll, or religious). In addition, the dataset includes three extra variables that reflect the commenter's gender, the category in which the remark was made, and the total number of answers for each comment.

B. DATA PRE-PROCESSING The collected data required preprocessing due to the presence of unstructured content remnants. In essence, it meant that the data needed to be cleaned up or trimmed to improve its correctness. Data cleansing, stop word removal, and tokenization were among the processes used to preprocess the data. We used a stop word filter to remove any unneeded terms from all text chats that corresponded to Bengali vocabulary. Stop words are those that do not supply sufficient information to establish which category a text belongs to. To make it easier to avoid distinguishing between capital and lowercase letters, we transformed all of the data to lowercase. To help with the feature extraction procedure, tokenization must also be applied to these text elements. Tokenization is one method for separating or isolating each word that accumulates in a text or discussion.

C. FEATURE EXTRACTION The preprocessed text conversation data will be turned into a vector space model, using Term Frequency Inverse Document Frequency (TFIDF) utilized to describe the text discussions using an extracted feature vector. TFIDF is primarily a method for quantifying or assessing a word's significance to a document or collection of documents. As a result, TFIDF's main distinguishing feature is its outstanding text performance and the acquisition of these terms' relative weights within sentences or documents. In addition to TFIDF, we will use word-level feature extraction techniques, known as the Bag of Words or "Bag of n Grams" representation. It implies that words are defined or represented solely by their appearance in a document, without regard for their location or order. The max df parameter, which is designed to exclude phrases that appear too frequently in the document, is one of several parameters used to integrate the vectorizer with a machine learning model.

D. CLASSIFICATION The recovered features are then loaded into an algorithm that trains and tests the classifier, assessing if it can detect cyberbullying. This concludes the categorization step of the proposed model. A wide range of machine learning techniques and algorithms will be used, including the Support Vector Machine (SVM), Random Forest, Logistic Regression (LR), and Passive Aggressive (PR) classifier. Each of these classifiers is evaluated using only a few assessment lattices. These criteria include the f-score, recall, accuracy, and precision.

ISSN: 2366-1313

5. CONCLUSION

We tried text classification techniques to detect cyberbullying in Bengali. Despite the fact that we've worked with a range of text-based classification algorithms, including SVM, Random Forest, Logistic Regression, and Passive Aggressive, more machine learning models or techniques, such as CNN and even NLP, can be applied to the dataset in the future.

REFERENCES

1. Abdhullah-Al-Mamun, Shahin Akhter, "Social media bullying detection using machine learning on Bangla text", 10th International Conference on Electrical and Computer Engineering (ICECE) 2018.
2. [Md Gulzar Hussain∗ , Tamim Al Mahmud (Member, IEEE), Waheda Akthar, "An Approach to Detect Abusive Bangla Text", International Conference on Innovation in Engineering and Technology (ICIET) 27-29 December, 2018
3. Md Faisal Ahmed, Zalish Mahmud, Zarin Tasnim Biash, Ahmed Ann Noor Ryen, Arman Hossain, Faisal Bin Ashraf R, "Bangla Text Dataset and Exploratory Analysis for Online Harassment Detection" Department of Computer Science and Engineering, Brac University, 4 th February, 2021
4. John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, Ammar Mohammed, "Social Media Cyberbullying Detection using Machine Learning", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019
5. Rohit Pawar, Rajeev R. Raje, "Multilingual Cyberbullying Detection System", IEEE International Conference on Electro Information Technology (EIT), 2019
6. B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying detection: A survey on multilingual techniques," in European Modelling Symposium (EMS), pp. 165–171, Nov 2016.
7. Michele Di Capua, Emanuel Di Nardo, and Alfredo Petrosino. "Unsupervised cyber bullying detection in social networks". In Pattern Recognition (ICPR), 2016 23rd International Conference on, pages 432–437. IEEE, 2016
8. Sani Muhamad Isa, Livia Ashianti, et al. "Cyberbullying classification using text mining". In Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on, pages 241–246. IEEE, 2017