International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Cyberbullying Detection using machine learning

Submitted by Logasree.S ,Harshini.M
B.Sc. Computer science
Arasu college of arts and science for women

Abstract—

Cyber bullying is the process of sending wrong messages to a person or community which causes heated debate with users. Cyber bullying is mostly seen in social networking sites where users reply to post with bullying words to threaten or insult other users. Cyber bullying is considered a misuse of technology. According to the latest survey done on all over the world data day by day, cases are increasing on cyber bullying. In order to solve this problem many natural language processing techniques are proposed by various authors which are time taking and not automatic. With the advancement of machine learning and artificial intelligence, models can be created and automatic detection can be implemented. To show this scenario live chat application is developed in python programming with multiple clients and one server and the Naive Bayes algorithm is used to train the model on a Twitter dataset and using this model live detection of cyber bullying is predicted and alert messages are shown on the chat application.

Keywords: Cyberbullying, Machine learning, Naïve Bayes, dataset.

PROBLEM STATEMENT:

Social networking and online chatting application provide a platform for any user to share knowledge and talent but few users take this platform to threaten users with cyberbullying attacks which cause issues in using these platforms.

OBJECTIVE:

To provide a better platform for users to share knowledge on social networking sites there is a need for an effective detection system that can automate the process of cyberbullying detections and take decisions.

I. INTRODUCTION

Social media is a platform that allows people to post any thing like photos, videos, documents extensively and interact with society . People connect with social media using their computers or smartphones. The most popular social media includes Facebook, Twitter, Instagram, tiktok and so on. Nowadays, social media is involved in different sectors like Education , business, and also for the noble cause. Social media is also enhancing the world's economy through creating many new job opportunities.

Although social media has a lot of benefits, it also has some drawbacks. Using this media, malevolent users conduct unethical and fraudulent acts to hurt others feelings and damage their reputation. Recently, cyberbullying has been one of the major social media issues. Cyberbullying or cyber-harassment refers to an electronic method of bullying or harassment. Cyberbullying and cyber-harassment are also known as online bullying. As the digital realm has grown and technology has progressed, cyberbullying has become relatively common, particularly amongst adolescents.

As the social lifestyle exceeds the physical barrier of human interaction and contains unregulated contact with strangers, it is necessary to analyse and study the context of cyberbullying. Cyberbullying makes the victim feel that he is being attacked everywhere as the internet is just a click away. It can have mental, physical, and emotional effects on the victim. Cyberbullying mainly takes place in the form of text or images on social media. If bullying text can be distinguished from non-bullying text, then a system can act accordingly. An efficient cyberbullying detection system can be useful for social media websites and other messaging applications to counter such attacks and reduce the number of cyberbullying cases. The objective of the cyberbullying detection system is to identify the cyberbullying text and also take its meaning into consideration.

II. RELATED STUDY

There are several works on machine learning-based cyber-bullying detection. A supervised machine learning algorithm was proposed using a bag-of-words approach to detect the sentiment and contextual features of a sentence . This algorithm shows barely 61.9% of accuracy. Massachusetts Institute of Technology conducted a project called Ruminate employing support vector machine to detect cyberbullying of YouTube comments. The researcher combined detection with common sense reasoning by adding social parameters. The result of this project was improved to 66.7% accuracy for applying probabilistic modelling. Reynolds proposed a language-based cyberbullying detection method which shows 78.5% of accuracy. The authors used the decision tree and instance-based trainer to achieve this accuracy. To improve cyberbullying detection, the author of the paper has used personalities, emotion and sentiment as the feature.

Several deep learning-based models were also introduced to detect the cyberbullying. Deep Neural Network-based model is applied for cyberbullying detection by using real-world data. The authors first analyse cyberbullying systematically then used transfer learning to do the detection task. Badjatiya has presented a method using deep neural net-work architectures for detecting hate speech. A convolutional neural network-based model has been proposed to detect cyberbullying. The authors employed word embedding where similar words have similar embedding. In a multi-modal context, Cheng research the novel issue of cyberbullying identification by collaboratively exploiting social media data. This challenge, however, is difficult due to the complex combination of both cross-modal associations among multiple methods and structural correlations between various social media sessions, and the complex attribute in-formation of different modalities. They propose XBully, a novel cyberbullying identification system to overcome these challenges, which first reformulates multi-modal social media data as a heterogeneous network and then tries to learn node embedding representations on it.

Many literatures on cyberbullying have concentrated on text analysis over the past few decades. Cyberbullying, however, is becoming multi-objective, multi-channel, and multi-form. The variety of bullying data on social platforms can not be met by conventional text analytical techniques.

Using Neural Networks to facilitate the identification of online bullying has become common in recent years. These Neural Networks are also based solely on or in conjunction with other layer types utilising Long-Short-Term-Memory layers. Buan introduced a new model for the Neural Network that can be applied in textual media to identify evidence of cyberbullying. The concept is made on existing architectures that merge the strength of Long-Short-Term-Memory layers with Convolutionary layers. In addition, their architecture features the use of stacked core layers, which demonstrates that their study enhances the Neural Network's efficiency. A new type of activation method is also included in the design, that is called ”Support Vector Machine like activation” By using L2 weight regularisation and a linear activation function in the activation layer along with using a Hinge loss function, the ”Support Vector Machine like activation” is accomplished.

Cyberbullying has recently been identified by users of online social networks as a significant national health problem and the creation of an effective detection model has considerable scientific merit. Al have introduced a collection of specific Twitter-derived features including behaviour, user, and tweet content. They have built a supervised machine learning solution for the detection of cyberbullying on Twitter based network. An assessment shows that, based on their proposed features, their HI established detection system obtained outcomes with a region under the receiver-operating characteristic curve of 0.943 and an f-measure of 0.936.

III. BULLYING DETECTION MODEL

In this section, we describe the cyberbullying detection framework which consists of two major parts as shown in 1. The first part is called NLP (Natural Language Processing) and the second part is named as ML (Machine learning). In the first phase, datasets containing bullying texts, messages or post are collected and prepared for the machine learning algorithms using natural language processing. The processed datasets are then used to train the machine learning algorithms for detecting any harassing or bullying message on social media including Facebook and Twitter.

A. Methodology

• Natural Language processing: The real world posts or text contain various unnecessary characters or text. For example, numbers or punctuation are irrelevant to bullying detection. Before applying the machine learning algorithms to the comments, we need to clean and prepared them for the detection phase. In this phase, various processing task including removal of all irrelevant characters like stop-words, punctuation and numbers, tokenizations, stemming etc. After the pre-processing, we prepare the two important features of the texts as follows:

1) Bag-of-Word: The machine learning algorithms can-not work directly with the raw text. So before applying the algorithms we must convert them to vector or numbers. So, the processed data is converted to Bag-of-Words (BoW) for the next phase.

2) TF-IDF: This is another features that we consider for our model. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure that can evaluate how relevant a word is to a document in a collection of documents. In bag of words, every word is given equal importance while in TF-IDF the words that occur more frequently should be given more importance as they are more useful for classification.

• Machine Learning: This module involves in applying various machine learning approaches like Decision Tree (DT), Random Forest, Support Vector Machine, Naive Bayes to detect the bullying message and text. The classifier with the highest accuracy is discovered for a particular public cyberbullying dataset. Next section, some common machine learning algorithms are discussed to detect cyberbullying from social media texts.

B. Machine Learning Algorithms

In this section, we discussed the basic mechanisms of several machine learning algorithms. We presented Decision Tree, Naive Bayes, Support Vector Machine in each subsection.

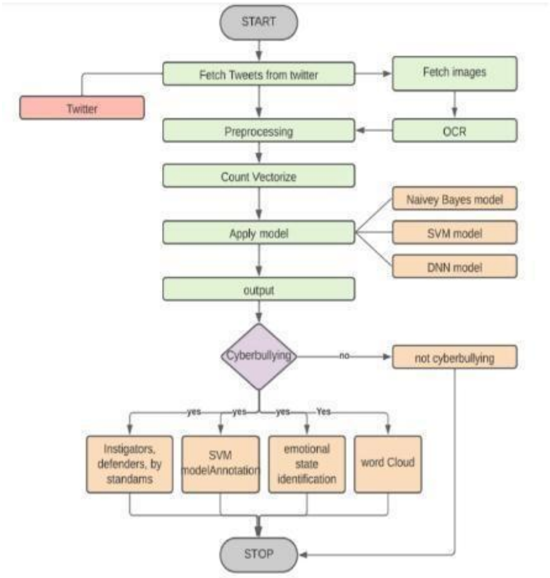
1) Decision Tree: The decision tree classifier can be used in both classification and regression. It can help represent the decision as well as make a decision. The decision tree is a tree- like structure where each internal node represents a condition, and each leaf node represents a decision. A classification tree returns the class where the target falls. A regression tree yields the predicted value for an addressed input.

2) Naive Bayes: Naive Bayes is an efficient machine learning algorithm based on Bayes theorem . The algorithm predicts depending on the probability of an object. The binary and multi-class classification problems can be quickly solved using this technique. Based on Bayes' Theorem it finds the probability of an event occurring given the probability of another event that has already occurred as follows:

$$\frac{p(y|X) = p(X|y) \times p(y)}{X}$$

3) Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning algorithm which can be applied in both classification and regression alike a decision tree. It can distinguish the classes uniquely in n dimensional space. Thus, SVM produces a more accurate result than other algorithms in less time. In practice, SVM constructs set a of hyper planes in a infinite-dimensional space and SVM is implemented with kernel which transforms an input data space into the required form. For example, Linear Kernel uses the normal dot product of any two instances as follows:

$$K(x, xi) = sum(x * xi)$$



IV. IV. FEATURES

- ❖ Detection of Non-Textual Cyberbullying

We are going to develop an application which has image in tweets or online data and we will fetch such image from twitter and after OCR classification will be done by our model SVM or naïve bayes model.

- ❖ Expanding Cyberbullying Role Detection beyond Victims and Bullies

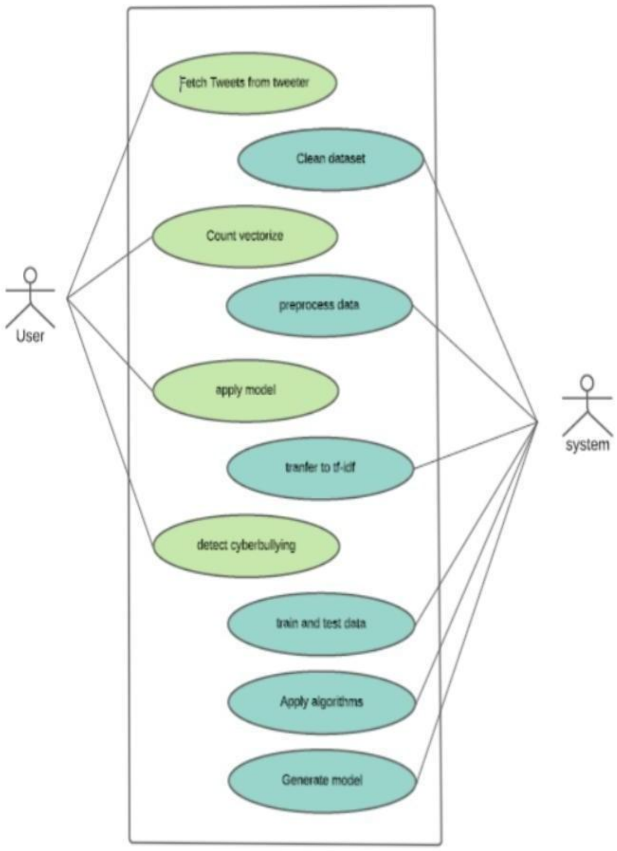
Roles such as instigators, defenders, and bystanders will be identified by us based on the algorithm model generated by us by collecting and labelling such type of data.

- ❖ Determining a Victim's Emotional State after a Cyberbullying Incident

A victim may change his/her profile details following such interactions, post content containing negative sentiments, or leave the network abruptly. Such instigating interaction can be flagged up for subsequent review by a human who can then follow-up with appropriate actions Twitter will not allow to go in the profile of user for this we might create our own system which can identify such changes and will determine how the bullying affected person.

- ❖ Word Representation Learning for Cyberbullying detection

correct psychometric categorization of the text. In future it is intended to improve the system developed by use more accurate dataset and to detect the cyberbullying or not. We also apply other machine learning algorithm and check the accuracy of models. Higher accuracy model will help to detect more accurate bullying. Another interesting direction for future work would be the detection of fine-grained cyberbullying categories such as threats, curses and expressions of racism and hate. When applied in a cascaded model, the system could find severe cases of cyberbullying with high precision. This would be particularly interesting for monitoring purposes. Additionally, our dataset allows for detection of participant roles typically involved in cyberbullying.



Experiments can be performed to generate word embedding’s from different datasets, ranging from general corpora (e.g., Wikipedia) to more specialised datasets (e.g. Abusive tweets) to compare their effectiveness for cyberbullying detection.

- Detecting Cyberbullying in Streaming Data and Real-time

We will determine the cyberbullying on twitter dataset oauthtoken will be generated on twitter account we will fetch the tweets.

- Evaluating Annotation Judgement

We will annotate the each twitter sentence and output will begenerated shown on text .

V. FUTURE MODIFICATION

The validity and accuracy of the predictive models to detect cyberbullying on twitter in this case primarily based on the

VI. CONCLUSION

The goal of this project is to the automatic detection of cyberbullying-related posts on social media. Given the information overload on the web, manual monitoring for cyberbullying has become unfeasible. Automatic detection of signals of cyberbullying would enhance moderation and allow to respond quickly when necessary. However, these posts could just as well indicate that cyberbullying is going on. The main aim of this project is that it presents a system to automatically detect signals of cyberbullying on social media, including different types of cyberbullying, covering posts from bullies, victims and bystanders.

VII. References

1. Cyberbullying Detection System on Twitter ieee paper
2. Methods for Detection of Cyberbullying: A Survey ieee paper
3. D. Poeter. (2011) Study: A Quarter of Parents Say Their Child Involved in Cyberbullying. pcmag.com. [Online].Available: http://www.pcmag.com/article2/0,2817,2388540,00.asp
4. J. W. Patchin and S. Hinduja, “Bullies move Beyond the Schoolyard; a Preliminary Look at Cyberbullying,” Yout Violence and Juvenile Justice, vol. 4, no. 2, pp. 148–169,2006
5. Anti Defamation League. (2011) Glossary of Cyberbullying Terms.adl.org.[Online].Available: http://www.adl.org/education/curriculum connections/cyberbullying/glossary.pdf
6. https://www.sciencedirect.com/topics/computer-Science/deep-neural-network
7. J. C. Platt, “Fast Training of Support Vector Machines using Sequential Minimal Optimization,” Advances in Kernel Methods, pp. 185–208, 1999. [Online]. Available: http://portal.acm.org/citation.cfm?id=299094.29915
8. K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the Detection of Textual Cyberbullying,” in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media (SWM’11), Barcelona, Spain, 2011.
9. Approaches to Automated Detection of Cyberbullying: A Survey ieee paper
10. https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.5627
11. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240924
12. https://1000projects.org/cyber-bullying-detection-using-machine-learning.html
13. https://www.slideshare.net/ashisharora965/detecting-the-presence-of-cyberbullying-using-computer-software
14. https://slideplayer.com/slide/11975230/
15. https://engineering.ucdenver.edu/current-students/capstone-expo/archived-expos/spring-2020/computer-science/csci14-cyberbullying-detection-system