Detection of Cyberbullying on social media Using Machine Learning

B. PRASANNA KUMAR[1], CH VISHNU VARDHAN REDDY[2], CH PHANENDRA REDDY[3], CH VENKATESH[4], G. MAHANVITH[5]

[1] *Associate Professor of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, AP, India.*

[2, 3, 4, 5] *B. Tech (CSE), KKR & KSR Institute of Technology and Sciences, Guntur, AP, India.*

Abstract*— Cyberbullying is a major problem encountered on the internet that affects teenagers and also adults. It has led to mishappenings like suicide and depression. Regulation of content on Social media platforms has become a growing need. The following study uses data from two different forms of cyberbullying, hate speech tweets from Twitter and comments based on personal attacks from Wikipedia forums to build a model based on the detection of Cyberbullying in text data using Natural Language Processing and Machine learning. Three methods for Feature extraction and four classifiers are studied to outline the best approach. For Tweet data, the model provides accuracies above 90% and for Wikipedia data, it gives accuracies above 80%.

***Indexed Terms*-- Cyberbullying, Hate speech, Personal attacks, Machine learning, Feature extraction, Twitter, Wikipedia**

I. INTRODUCTION

Now more than ever technology has become an integral part of our life. With the evolution of the internet. Social media is trending these days. But like all the other things misusers will pop out sometimes late sometimes early but there will be for sure. Now Cyberbullying is common these days.

Sites for social networking are excellent tools for communication among individuals. The use of social networking has become widespread over the years, though, in general people find immoral and unethical ways of negative stuff. We see this happening between teens or sometimes between young adults. One of the negative stuff they do is bully each other over the internet. In an online environment, we cannot easily say whether someone is saying something just for fun or if there may be another intention of him. Often, with just a joke, "or don't take it so seriously," they'll laugh it off Cyberbullying is the use of technology to harass, threaten, embarrass, or target another person. Often this internet fight results in real-life threats for some individuals. Some people have turned to suicide. It is necessary to stop such activities at the beginning. Any actions could be taken to avoid this for example if an individual's tweet/post is found offensive then maybe his/her account can be terminated or suspended for a particular period.

II. LITERATURE REVIEW

A lot of research has been done to find possible solutions to detect Cyberbullying on social networking sites. Hsien used an approach using keyword matching, opinion mining and social network analysis and got a precision of 0.79 and a recall of 0.71 from datasets from four websites. Patxi Gal´an-Garc´ıa et al. proposed a hypothesis that a troll(one who cyberbullies) on a social networking site 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 authors. Later it was used in a Case Study in a school in Spain where out of some suspected students of 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 behaviours 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. suggested a B-LSTM technique based on concentration. Banerjee et al. used KNN with new embedding's to get a precision of 93%. Kelly Reynolds, April Kontostathis and Lynne Edwards propose a Forming(A forum for anonymous questions answers) dataset which gives a 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 an 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 the Youtube dataset for the same and achieved an F1 score of 0.97 using the Bidirectional Long Short-Term Memory(BLSTM) model.

Sweta Agrawal and Amit Awekar used similar same datasets for training Deep Neural Networks but one of its key focus is swearing 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 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 the probability of being bullied.

III. PROPOSED SYSTEM & RESULTS

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.
Tokenization: In tokenization, we split the raw text into meaningful words or tokens. For example, the text "we will do it" can be tokenized into 'we', 'will', 'do', 'it'. Tokenization can be done in 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 Eg for the regular expression '\w+' all the alphanumeric tokens are extracted.

Stemming: Stemming is the process of converting a word into a root word or stem. Eg for three words 'eating' 'eats' 'eaten' the stem is 'eat'. Since all three branch words of root 'eat' represent the same thing it should be recognized as similar. NLTK offers 4 types of stemmers: Porter Stemmer, Lancaster Stemmer, Snowball Stemmer and Regexp Stemmer. The following project uses PorterStemmer.

Stop word Removal: Stop words are words that do not add any meaning to a sentence eg. 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 which 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 to improve.

Remote User:
In this module, there are n numbers of users are present. Users should register before doing any operations. Once user registers, their details will be stored in the database. After registration is successful, he has to log in by using an authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT CYBERBULLYING, and VIEW YOUR PROFILE.

Service Provider:
In this module, the Service Provider has to log in by using a valid user name and password. After login successful he can do some operations such as Login,

Train and Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Cyberbullying Predict Type Details, Find Cyberbullying Prediction Ratio on Data Sets, Download Cyber Bullying Prediction Data Sets, View Cyberbullying Prediction Ratio Results, View All Remote Users.

Algorithm:

Step 1: Start

Step 2: Login using the registered user credentials, if not, then register.

Step 3: Click on view profile to see full profile details.

Step 4: Click on Predict button after entering the message to see if the text is offensive & cyberbullying or not.

Step 5: On clicking the Predict button three ML models (SVM, Logistic Regression & Naive Bayes) which are already trained with the training data set will run the given message text and predict whether it is offensive or not.

Step 6: After the prediction is completed the model result with the highest accuracy will be shown on the display.

Step 7: Stop

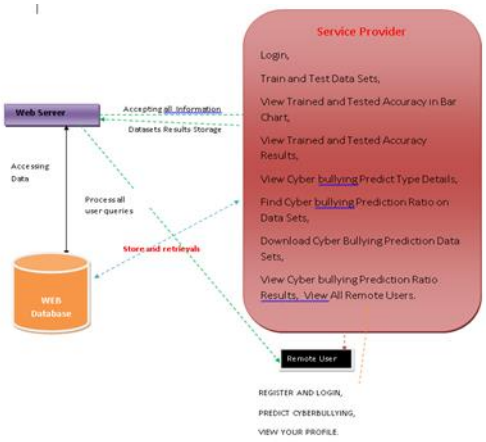
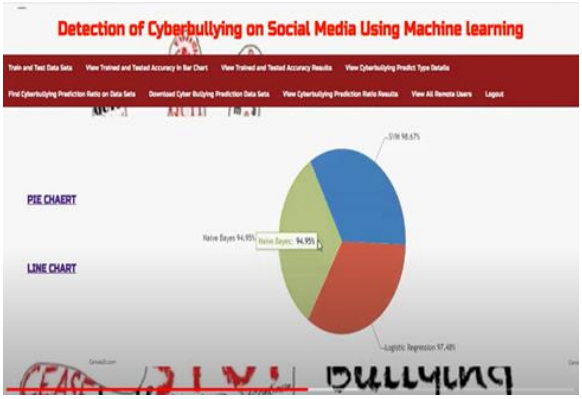


Figure 1: Architecture of the proposed system

Experimental Results:

User Registration

Prediction:



CONCLUSION

Cyberbullying across the internet is dangerous and leads to mishappenings like suicides, depression etc. therefore there is a need to control its spread. Therefore, cyberbullying detection is vital on social media platforms. With the availability of more data and better-classified user information for various other

forms of cyberattacks 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 the detection of cyberbullying 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% 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 the context of features proved effective in both datasets giving similar results in comparatively fewer features when combined with Multi-Layered Perceptrons. As seen by changing nature.

REFERENCES

[1] R. Kessler, E. Bromet, P. Jonge, V. Shahly, and Marsha., "The burden of depressive illness," Public Health Perspectives on Depressive Disorders, 2017.

[2] W. H. Organisation, "Mental health: Fact sheet," https://www.euro.who.int/en/health-topics/noncommunicablediseases/ mental-health, 2019.

[3] S. Guntuku, D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," Current Opinion in Behavioral Sciences, 2017.

[4] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: A content analysis in heidelberg," Biomed Inform Insights, 2010.

[5] Preotiuc-Pietro, D., Sap, M., Schwartz, A., Ungar, L.: Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. In Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, North American Chapter of the Association for Computational Linguistics. (2015)