CYBER BULLYING DETECTION IN TWITTER & SOCIAL MEDIA

[1]M DAMODHAR, [2]POLOJU VIKRAM, [3]PEDDATHUMBALAM ANIL KUMAR,
[4]MEKADONA ASHOK, [5]A. EMMANUEL RAJU

[1234]B.Tech Student, [5]Assistant Professor

Department of CSE

Dr. K. V. Subba Reddy Institute Of Technology, Kurnool

ABSTRACT

Cyberbullying is a major problem encountered on internet that affects teenagers and also adults. It has lead to mis happenings 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 Twittter and comments based on personal attacks from Wikipedia forums to build a model based on 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%.

1. 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 as all the other things mis users will pop out sometimes late sometime early but there will be for sure. Now Cyber bullying is common these days.

Sites for social networking are excellent tools for communication within individuals. 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 stuffs they do is bullying each other over the internet. In online environment we cannot easily said that whether someone is saying something just for fun or there may be other intention of him. Often, with just a joke, "or don't take it so seriously," they'll laugh it off Cyber bullying is the use of technology to harass, threaten, embarrass, or target another person. Often this internet fight results into real life threats for some individual. 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.

So, what is cyber bullying??

Cyber bullying is harassment, threatening, embarrassing or targeting someone for the purpose of having fun or even by well-planned means.

2. SYSTEM ANALYSIS

Architecture Diagram
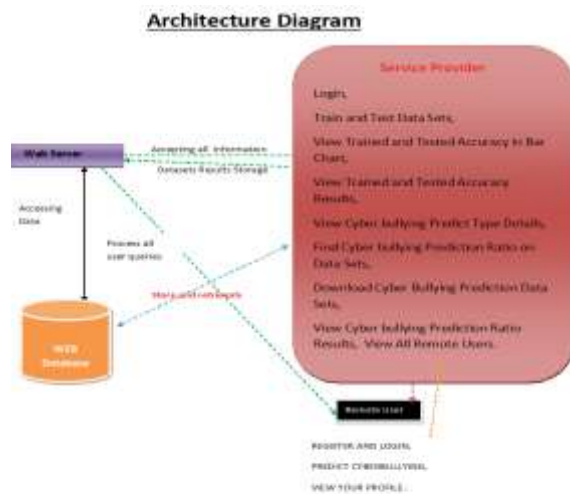
EXISTING SYSTEM

❖ Hsien[1] 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.[2] 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. [3] 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.[4] suggested a B-LSTM technique based on concentration. Banerjee et al.[5]. used KNN with new embeddings to get an 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

➢ Tf-Idf 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

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 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 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 rule which in the 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 Porter Stemmer.

Stop word Removal: Stop words are words that do not add any meaning to a sentence eg. some stop words for 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 and helps in improving performance.

Advantages
  ➢ Common Bag of Words model takes as input of multiple words and predicts the word based on the context. Input can be one word or multiple words.
  ➢ CBOW model takes a mean of context of input words but two semantics can be clicked for a single word. i.e. two vector of Apple can be predicted. First is for the firm Apple and next is Apple as a fruit.

3. MODULES

Service Provider
In this module, the Service Provider has to login by using 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 Cyber bullying Predict Type Details, Find Cyber bullying Prediction Ratio on Data Sets, Download Cyber Bullying Prediction Data Sets, View Cyber bullying Prediction Ratio Results, View All Remote Users.

View and Authorize Users
In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

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

4. CONCLUSIONS

Cyber bullying across internet is dangerous and leads to mis happenings 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 availability of more data and better classified user information for various other forms of cyber attacks Cyber bullying 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] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in Proceedings of 4th
International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.

[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.

[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," 2011, doi: 10.1109/ICMLA.2011.152.

[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.

[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.

[9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv.2018.

[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," 2016, doi: 10.1109/ASONAM.2016.7752420.

[11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," 2016, doi: 10.18653/v1/n16-2013.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," 2017.

[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," 2017, doi: 10.1145/3038912.3052591.

[14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif. Intell. Rev., vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.