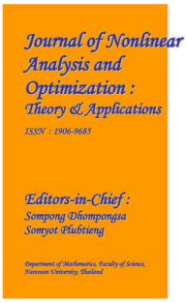
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.15 : 2024
ISSN : **1906-9685**



DETECTION OF CYBERBULLING ON SOCIAL MEDIA USING MACHINE LEARNING

[1]*K.Divya,*[2]*Amgoth Naresh,*[3]*R. Dheeraj*
[1,2,3]*Assistant Professor*
Department of CSE
Visvesvaraya College Of Engineering & Technology,Ibrahimpatnam,Telangana

ABSTRACT
Cyberbullying is an activity of sending threatening messages to insult person. To prevent cyber victimization from the activity is challenging. Cyberbullying is a major problem encountered on internet that affects teenagers and also adults. It has lead to mishappening like suicide and depression. Cyberbullying detection is very important because the online information is too large so it is not possible to be tracked by humans. 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 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%.

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 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

1.1 BACKGROUND

Researches on Cyber bullying Incidents show that 11.4% of 720 young peoples surveyed in the NCT DELHI were victims of cyber bullying in a 2018 survey by Child Right and You, an NGO in India, and almost half of them did not even mention it to their teachers, parents or guardians. 22.8% aged 13-18 who used the internet for around 3 hours a day were vulnerable to Cyber bullying while 28% of people who use internet

more than 4 hours a day were victims. There are so many other reports suggested us that the impact of Cyber bullying is affecting badly the peoples and children between age of 13 to 20 face so many difficulties in terms of health, mental fitness and their decision making capability in any work. Researchers suggest that every country should have to take this matter seriously and try to find solution. In 2016 an incident called Blue Whale Challenge led to lots of child suicides in Russia and other countries . It was a game that spread over different social networks and it was a relationship between an administrator and a participant. For fifty days certain tasks are given to participants . Initially they are easy like waking up at 4:30 AM or watching a horror movie . But later they escalated to self harm which let to suicides. The administrators were found later to be children between ages 12-14.

II. LITERATURE SURVEY

Lot of research have been done to find possible solutions to detect Cyberbullying on social networking sites.Ting,Ito 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 recall of 0.71 from datasets from four websites.PatxiGal´an- Garc´ıa et al.

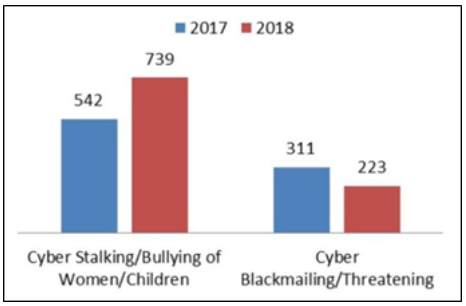


Fig.1. Cyberbullying cases in India 2017-2018

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 doesn't have a real account to fool such systems or experts who can change writing styles and behaviour so that no patterns are found. For changing writing styles more efficient algorithms will be needed Mangaonkaretal.

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. Zhouetal. Suggested a B-LSTM technique based on concentration. Banerjeeetal. Used KNN with new embeddings to get an precision of 93%. Kelly Reynolds, April Kontostathis and Lynne Edwards.

Propose a Formpring(A forum for anonymous questionsanswers) 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. 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. 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.

III. SYSTEM ANALYSIS

3.1 PROPOSED METHODOLOGY

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. Fig. 2 describes the methodology used for solving the problem which is applied on both the datasets.

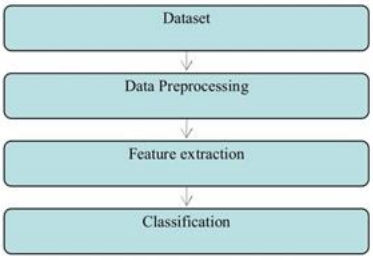


Fig.2. Methodology

3.1 Modules

3.1.1 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.

3.1.2 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.

3.1.3 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.

IV. OUTPUT SCREENS

Fig.3. Home Page

Admin pages

Fig.4. Service provider login page

Fig.5. Service provider home page

Fig.6. Train and Test datasets view

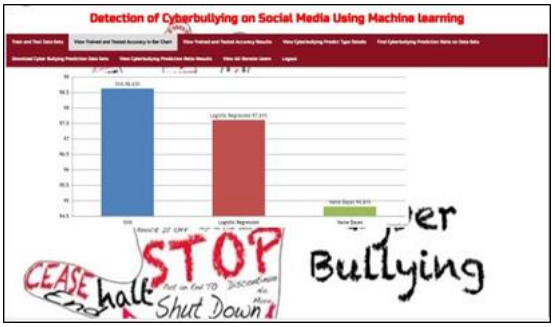


Fig.7. Trained and Tested Accuracy in Bar Chart

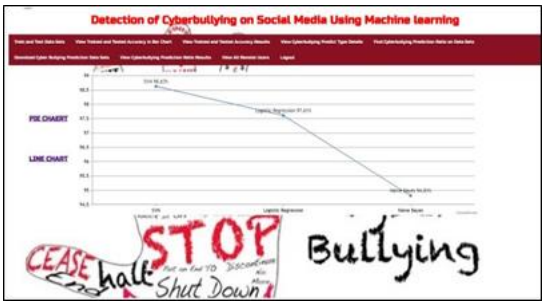


Fig.8. Trained and Tested Accuracy Results

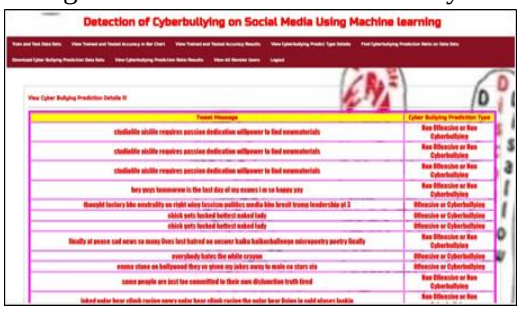


Fig.9. View Cyberbullying Predict Type Details

Fig.10. Find Cyberbullying Predict Ratio on DataSets

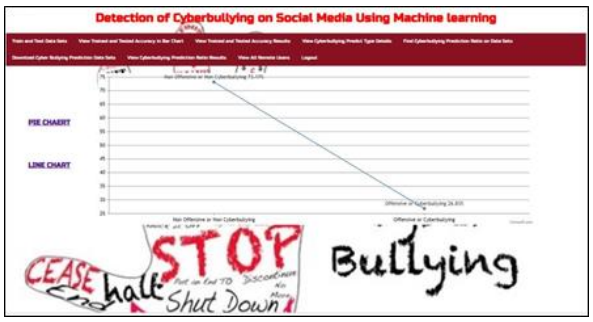


Fig.11. View Cyberbullying Predict Ratio Results

USER PAGES

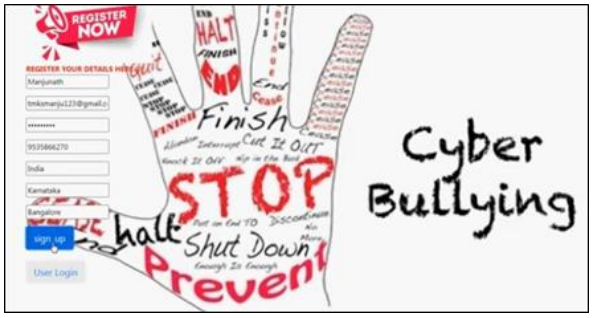


Fig.12. Registration page

Fig.13. User login page

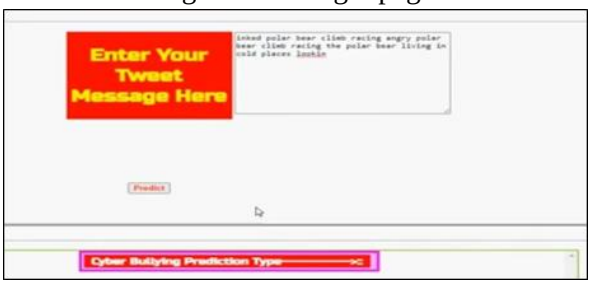


Fig.14. User home page

Fig.15. Prediction

V. 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 availability 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.

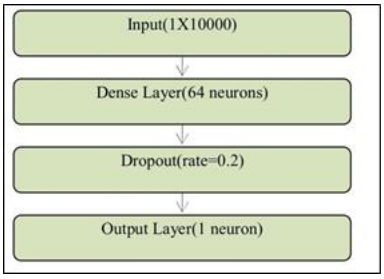


Fig.16. Multi Layered Perceptron used for Bag of Words and TF-IDF inputs

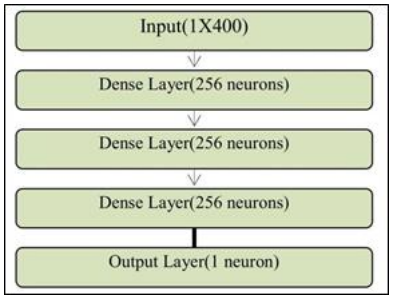


Fig.17. Multi Layered Perceptron used for Word2Vec input

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.

The overall aim of the project "cyberbullying detection using machine learning" is to develop a system that automatically classifies comments and messages as bullying or non-bullying and also remove the bullying comments from the web application

FUTURE SCOPE

Complex problems like cyberbullying, which have various problems embedded in, are difficult to trace with the normal system. Especially, image-based social cyberbullying post-detection is a challenging task. This research explored deep learning and transfer learning frameworks to find the best-suited model to predict image-based cyberbullying posts on social platforms. The deep learning-based 2DCNN has initially experimented and, by tuning their hyperparameters, achieved the accuracy value of 69.60%. On the other hand, the transfer learning models VGG16 and InceptoionV3 always achieved better prediction accuracy. The VGG16 achieved an accuracy value of 86% whereas, InceptionV3 achieved 89% accuracy. Hence, the transfer learning models VGG16 and InceptoionV3 have an accuracy margin of 16.40 and 19.40%, respectively, compared to the best configured 2DCNN model. Therefore, it can be concluded that the proposed system detects most of the image-based cyberbullying posts.

REFERENCE

1. Smith PK, Mahdavi J, Carvalho M, Fisher S, Russell S, Tippett N (2008) Cyberbullying: its nature and impact in secondary school pupils. J Child Psychol Psychiatry 49(4):376–385
2. Ak Şerife, Özdemir Y, Kuzucu Y (2015) Cybervictimization and cyberbullying: the mediating role of anger, don't anger me! Comput Human Behav 49:437–443
3. Kumari K, Singh JP, Dwivedi YK, Rana NP (2020) Towards cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Comput 24(15):11059–11070
4. 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.
5. A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.
6. R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.
7. https://ceoworld.biz/2018/10/29/countries-where-cyber-bullying-was-reported-the-most- in-2018/.
8. https://keras.io/api/applications/.
9. Keras.applications.inception_v3.preprocess_input.