iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.45847

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

A Review Paper on Cyber Harassment Detection Using Machine Learning Algorithm on Social Networking Website

Miss. Sneha Gajanan Sambare[1], Dr. Sanjay. L. Haridas[2]
[1]*PG Student, (M. Tech. Electronics Engineering), JD College of Engineering & Management*
[2]*Professor, Department of Electronics & Telecommunication Engineering, JD College of Engineering & Management*

Abstract: ***Cyber bullying is nothing but bullying with the use of digital technology. Cyber bullying can take place on social media platforms, messaging platforms, gaming platforms and mobile phones etc. Cyber bullying is a big issue that is encountered by the individual on internet that affects teenagers as well as adults. It has lead to nuisances like suicide, mental health problems and depression. Therefore regulation of content on Social media platforms has become a growing need. Also detecting Cyber bullying detection at early stages can help to alleviate impacts on the victims. In the following research we make use of data from two different data sets. The first one is tweets from Twitter and second one is comments based on personal attacks from Wikipedia forums. The approach is to build a model based on detection of cyber bullying in text data using Natural Language Processing and Machine learning. We try to build a model which will provide accuracy up to 90% for tweets and accuracy up to 80% for Wikipedia forums. Cyber bullying detection will be done as a binary classification problem where we are detecting two major form of cyber bullying: hate speech on Twitter and Personal attacks on Wikipedia and classifying them as containing Cyber bullying or not. The end result of this proposed work may reduce negative impacts on the victims as a result of early detection.***
Keywords: Cyber bullying, Social Media, Machine Learning,

I. INTRODUCTION

With the development of internet, social media is trending these days. Everyone jokes with everyone. But the problem arises when someone or a group of people starts to joke at you instead of joking with you. More often than not online abuse turnout to be real life threat to the victim. Cyber bullying is repeated behavior which aimed at scaring, angering or shaming an individual even after being asked to stop by the aimed individual. Cyber bullying is common these days and can hurt the victim mentally, emotionally or physically or all of them at once.

The feeling of being targeted can lead to people from sharing their problems and dealing with the problems. The victim can also turn towards drugs alcohol to deal with their mental and physical pain. Bullying via textual messages or pictures or videos on social media platforms has proven to be very harmful for adults. The term Bullying can be defined as an aggressive, intentional act that is carried out by a group or an individual against a victim who cannot easily defend him or herself repeatedly and over time even after being asked to stop.

The problems associated with cyber bullying makes it compulsory to detect it before it turns out t be a crime. Since until cyber bullying is detected it can't be stopped. Hence for bullying to stop, it needs to be identified first and then it must be reported. Early detection will also helps in supporting the victim and to help him dealing with the problems arising due to bullying. In this project we will be building a model that will be composed of two data sets one from twitter data set "Hate Speech Twitter Dataset by Waseem, Zeerak and Hovy" that contains 17000 tweets labeled for sexism or racism and second one from Wikipedia Dataset The Wikipedia dataset by Wulczyn, Thain and Dixon that contains 1M comments labeled for Personal attacks. For the analysis 40000 comments are taken into consideration from the dataset from which 13000 comments are labeled as Cyber bullying due to personal attack. After collecting data set we will be training the system in order to detect bullying.

This detection will be done as a binary classification problem. The algorithm used will be Natural Language Processing and Machine learning. As soon as bullying is detected it needs to be reported in order to stop it. The end result will implicate positive results on the victim's mental and physical issues.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

II. BACKGROUND

Researches on Cyber bullying incidents show that 11.4% of 720 young people 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 report it to their teachers, parents or guardians as they were so terrified. 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 of cyber bullying. There are so many other reports suggested us that the impact of cyber bullying is affecting badly the people and children between ages of 13 to 20. People are facing so many difficulties in terms of health, mental fitness and their decision making ability in work and other areas of life. Researchers also suggest taking cyber bullying seriously on the national level and try to find solution to cyber bullying. Cyber bullying should be considered as national level issue. As it is known to most of us that in the year 2016 an incident called Blue Whale Challenge led to lots of child suicides in known as Blue Whale Challenge arises in Russia and other countries. The dangerous game that spread over different social networking websites and it was a one to one relationship between an administrator and a participant. For fifty days specific tasks were given to participants starting from easy ones and then leading to the fatal events. Initially they were easy like waking up at 4:30 AM or watching a horror movie and then it finally ends up in suicide attempts. Hence it is quite necessary that these kinds of incidents must be reported at prior stages so as to avoid the hazardous effect.

III. MOTIVATION

1) To eradicate the evils existing on social media with the use of proper technology.
2) To help in identifying the crimes on social media and also identifying the criminals before the situation aggravates
3) Use of machine learning algorithms to detect the bullying on social media.
4) Existing methods are not proven to be very efficient methods of detection of cyber bullying so a system is must for countering cyber bullying.

IV. PROBLEM STATEMENT

1) Cyber bullying is a new type of bullying that follows an individual from the entrance of their work place and back to their homes.
2) Individuals are bullied from the start of the morning till they reach their homes at night and go to bed for sleep.
3) Existing methods are applied only when bullying becomes an act of crime not before that.
4) So it is very necessary to detect and monitor the anti-social elements before they do a crime.

V. LITERATURE REVIEW

1) Manuel F. López-Vizcaíno, Francisco J. Nóvoa, Victor Carneiro and Fidel Cacheda [16] proposed a two machine model which was a learning model for early detection of cyberbullying. They did the experiments using real world data set and following a time aware evaluation.
2) Syed Mahbub, Eric Pardede and A. S. M. Kayes [15] proposes analysis of the effects of predatory approach words in the detection of cyberbullying and proposes a mechanism for generation of a dictionary of such approach words. This was systematic approach design that takes the generation of keyword dictionary into consideration.
3) Patxi Galan-Garcia [2] proposed a hypothesis that a troll (cyber bully) on social networking sites under a fake profile always has a real profile to check how others view the fake profile. He proposed a machine learning approach to determine such profiles. The identification process studied some profiles that have some kind of close relationship to them. The method used was to use ML to select profiles to study, retrieve information about tweets, select features to be used from profiles, and find the author of the tweets. 1900 tweets from 19 different profiles were used. It had an accuracy of 68% for author identification. This was later used in a case study at a school in Spain, where cyberbullying had to find the real owner of the profile among some of the suspected students, and this method worked in the case. The following method still has some drawbacks. For example a case where the trolling account doesn't have a real account to fool the system or experts who can change the writing style and behavior so that no patterns are found. Changing writing style will require more efficient algorithms.
4) Kelly Reynolds, April Kontostathis and Lynne Edwards [6] proposed a Formspring (A forum for anonymous question answers) dataset which gives recall of 78.5% accuracy using Machine learning Algorithms and oversampling due to imbalance in cyber bullying posts. They used the labeled data, in conjunction with machine learning techniques which utilizes Weka tool kit, to train a computer to recognize harassment contents.
5) Maral Dadvar and Kai Eckert [8] trained deep neural networks on Twitter, Wikipedia, Formspring datasets and used the model on YouTube dataset and investigated the performance of their model in new social media platforms for the same and achieved F1 score of 0.97 using Bidirectional Long Short-Term Memory(BLSTM) model.

6) Sweta Agrawal and Amit Awekar [9] used similar same datasets for training Deep Neural Networks they provided several useful insights about cyberbullying detection. Their proposed system was basically the first of its kind that systematically analyzes cyberbullying detection using deep learning based models and transfer learning. They resolved that how the vocabulary for such models changes across various Social Media Platforms.

VI. PROPOSED SYSTEM

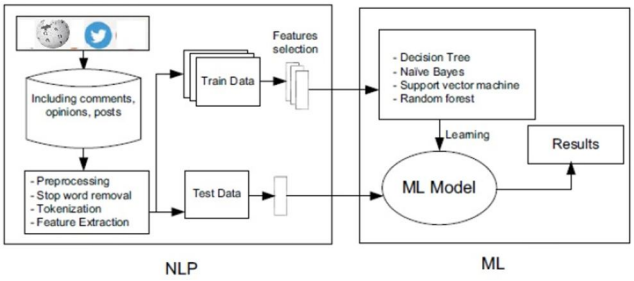


Figure: Proposed System

The proposed system consists of two datasets that we will be using to train our model.

A. Twitter Dataset: The Twitter Dataset is combined from two datasets containing hate speech: Hate Speech Twitter Dataset by Waseem, Zeerak and Hovy, Dirk [11] which contains 17000 tweets labeled for sexism or racism.

The tweets are mined using the annotations. 5900 tweets are lost due to accounts being deactivated or tweet deleted and second dataset of Hate Speech Language Dataset by Thomas Davidson, Dana Warmsley, Michael Macy and Ingmar Weber, [12]. It contained 25000 tweets obtained by crowd sourcing. This gives total 35787 tweets for the task distribution. For the following dataset, 70 % (25,050) of this dataset is used as training data and 30 percent as testing data (10,737).

B.Wikipedia Dataset The Wikipedia dataset by Wulczyn, Thain and Dixon [13] contains 1M comments labeled for Personal attacks. For our research analysis 40000 comments are used from the databanks from which 13000 comments are labeled as cyber bullying due to smear campaign. These comments are extracted from conversations between editors of pages on Wikipedia labeled by 10 annotators via Crowd Flower. For this dataset the same split (70 percent i.e 28000 to training data and 30 percent i.e 12000 to testing data) is used. Most of the existing system techniques are manual, which includes human intervention and decision making. Also existing system looks for patterns that already exist in the data. To overcome these deficiencies we are making use of Support Vector Machine (SVM). SVM is a computer algorithm that assigns the labels to object based on examples. For instance, an SVM can learn to categorize fraudulent or no fraudulent credit card activities by examining hundreds or thousands of fraudulent and or no fraudulent credit card activity reports. SVM is basically used to plot a hyper plane that creates a boundary between data points in number of features (N)-dimensional space. Linear SVM is used in the following case which is optimum for linearly separable data. In case of 0 misclassification, i.e. the class of data point is accurately predicted by our model, we only have to change the gradient from the regularization arguments. A random forest consists of different individual decision trees which individually predict a class for given query points and the class with maximum votes will be the final result. Decision Tree is a building block for random forest which provides a prediction by decision rules learned from feature vectors. A group of these uncorrelated trees provides a more corrective decision for classification. First we will design a Python language based portal which will in which we will be training our system using the algorithm with the help of data sets. Python is highly readable language also it has fewer syntactical constructions than other languages. Using Python language based portal, the two data sets and SVM we will train our network and as the system is trained if we enter the keywords or group of words then the system will show whether the entered content is bullying or not. A flag or warning message will pop out to show whether the entered content is bullying or not.

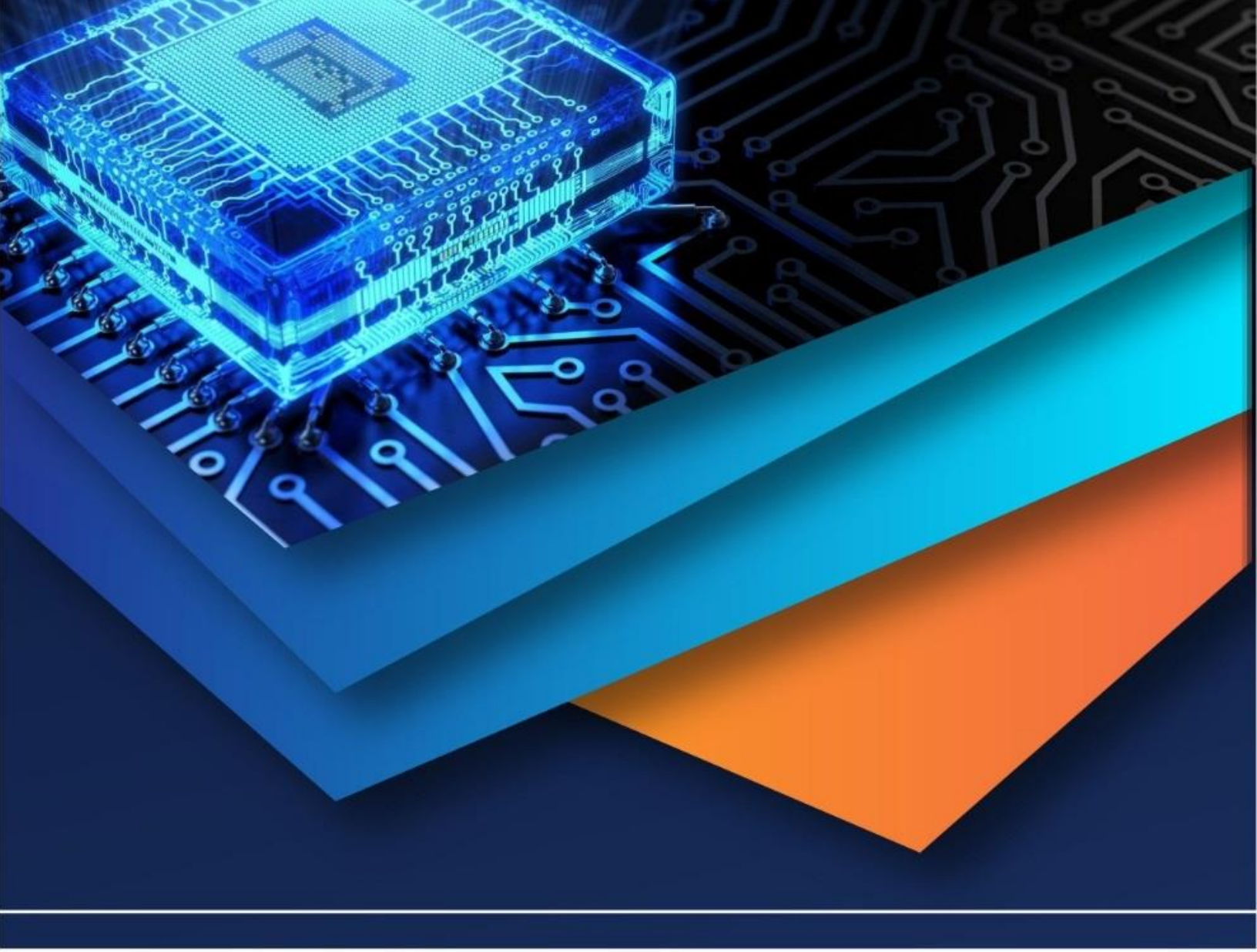
VII. CONCLUSION

A larger part of cyberbullying detection aimed at routine detection of cyberbullying incidents on online platforms focuses on machine learning models to detect bullying. Due to ever changing nature of online platforms detection of such events can have multiple ways. Cyber bullying on the online platforms is very menacing and leads to unfortunate events like depression, mental and physical illness and even suicide. Hence there is a need to control its scattering.

This leads to important aspect of online activities that is cyber bullying detection on social media platforms. In this proposed system we presented an idea for detection of cyber bullying and if integrated with real time applications it will help in restricting abusers bullying. We demonstrated the build up for two types of data: Hate speech Data on Twitter and Personal attacks on Wikipedia. For Hate speech Natural Language Processing techniques proved effective having accuracy of over 90% using basic Machine learning algorithms. As a result it gives better results with BoW and TF-IDF models instead of Word2Vec models. However, personal attacks were difficult to detect through the model because of the fact that the comments usually did not use any common sentiment that could be used to train the network. Word2Vec models which utilizes context of features proved effective in both datasets giving like to like results in comparatively less features after combining with Multi Layered Perceptrons. Additional perspective can also be taken into consideration for more specific results in future works. The nature of attacks are required to be studied further for a more specific and accurate detection.

REFERENCES

[1] 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 SocioCultural Computing, BESC 2017, 2017.
[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.
[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, “Collaborative detection of cyberbullying behavior in Twitter data,” 2015.
[4] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” 2016.
[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, “Detection of Cyberbullying Using Deep Neural Network,” 2019.
[6] K. Reynolds, A. Kontostathis, and L. Edwards, “Using machine learning to detect cyberbullying,” 2011.
[7] J. Yadav, D. Kumar, and D. Chauhan, “Cyberbullying Detection using Pre-Trained BERT Model,” 2020.
[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 cyber bullying in social networking sites,” 2016.
[11] Z. Waseem and D. Hovy, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” 2016.
[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.
[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] “Detection of Harassment Type of Cyberbullying: A Dictionary of Approach Words and Its Impact” Syed Mahbub, Eric Pardede and A. S. M. Kayes
[16] “Early detection of cyberbullying on social media networks” Manuel F. López-Vizcaíno, Francisco J. Nóvoa, Victor Carneiro and Fidel Cacheda.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY