ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

Detection of Cyber bullying on Social Media Using Machine Learning

[1]MR.RAMA BHADRA RAO MADDU, [2]DANIEL JOQUIS EDA

[1](Assistant Professor), MCA, Swarnandra College

[2]MCA, scholar, Swarnandra College

ABSTRACT

Teens and adults alike are not immune to the pervasive online scourge of cyberbullying.Problems like sadness and suicide have resulted from it. The need for content regulation on social media platforms is on the rise. This research takes a look at two types of cyberbullying—hate speech tweets from Twittter and comments based on personal attacks from Wikipedia forums—to create a model for detecting cyberbullying in text data using NLP and ML. In order to determine the optimal strategy, three feature extraction techniques and four classifiers were investigated. The model's accuracy for Tweet data is above 90% and for Wikipedia data it's over 80%. When it comes to interpersonal connection, social networking sites are second to none. Despite the fact that social media use has skyrocketed over the years, many still engage in harmful, immoral behavior on these platforms. Sometimes this occurs between young adults and other times it occurs between teenagers. One of the worst things they do is engage in cyberbullying. We have no way of knowing whether someone is speaking online for pleasure or if there is some ulterior motive behind their words.

1.INTRODUCTION

More than ever before, technology is intrinsic to how we live our lives. Thanks to the development of the internet. Trending topics in social media right now. The same is true with misusers; they may appear late or early, but they will always be there. Today, cyberbullying is a prevalent problem. Social networking sites are great for people to communicate with one other. Although more and more individuals are using social media, many still post nasty, immoral, and unethical content. When this occurs, it is usually between teenagers or young adults. They engage in harmful activities such as cyberbullying. Whether someone is saying something for fun or has ulterior motives, it's hard to tell in an online setting. Saying something like, "or don't take it so seriously," usually gets people to laugh it off. A person is being bullied or targeted when they utilize technology to threaten, shame, or harass another individual. Threats to people's physical safety are a common outcome of these online disputes. A number of individuals have attempted suicide. At the outset, it is essential to halt such actions. If someone's tweet or post is deemed objectionable, for instance, it may be possible to delete or suspend their account for a certain amount of time in order to prevent this. What exactly is cyberbullying, then?

Whether done in jest or with malice, cyberbullying may take many forms, including but not limited to threats, embarrassment, and harassment. Literature Review on Cyberbullying In 2018, a poll conducted by the Indian non-governmental organization Child Right and You found that 11.4% of the 720 youths questioned in the NCT DELHI had experienced cyberbullying. Almost half of those affected did not even inform their instructors, parents, or guardians about the incident. Internet users between the ages of 13 and 18 who spent three hours per day online were almost twice as likely to be victims of cyberbullying as those who spent more than four hours per day online. One must consider

Children and young adults (those between the ages of 13 and20) have

significant challenges related to their physical and mental well-being as well as their capacity to make sound decisions in all areas of life, according to several publications. Scientists believe that all nations should investigate this issue thoroughly. Lots of kids in Russia and other countries committed suicide in 2016 after an event known as Blue Whale Challenge. A connection between a game's administrator and a player blossomed as the game expanded across several social networks. Over the course of fifty days, participants are assigned specific tasks to complete. Like getting up at 4:30 in the morning or seeing a scary movie, they're simple at first. They started off little, but eventually progressed to self-harm, which ultimately led to suicides. It was only afterwards that it was discovered that the administrators were really 12–14 year olds.

2.LITERATURE SURVEY

The topic of cyberbullying and its detection on social media platforms has been the subject of much study. By combining keyword matching, opinion mining, and social network analysis, Ting,IFig Hsien[1] was able to get a recall of 0.71 and a precision of 0.79 utilizing datasets from four different websites.According to the theory put forward by Patxi Gal'an-Garc'ıa et al. [2], cyberbullies who use social media platforms often maintain a genuine profile alongside their phony one, in order to gauge how others perceive it. To find these profiles, they suggested using a machine learning method. As part of the identifying procedure, we looked at profiles that are similar to them. The procedure included picking profiles to analyze, collecting data from tweets, choosing attributes to utilize from profiles, and finally, using ML to identify the tweet's source. A total of 1900 tweets from 19 distinct accounts were analyzed. When it came to author identification, it was 68% accurate. Afterwards, it was used in a Case Study at a Spanish school, where the system was successful in identifying the true owner of a profile among many pupils who were suspected of cyberbullying. A few problems remain with the following approach. Such programs or experts may

modify writing styles and behaviors such that no patterns are identified, for instance, in a scenario when the trolling account does not have an actual account. More effective algorithms are required for modifying writing styles. In their collaborative detection technique, Mangaonkar et al. [3] suggested using a network of interconnected detection nodes, each of which might utilize a unique or shared algorithm to generate findings. The authors are P. Zhou and colleagues. A 93% success rate was achieved by Banerjee et al.[4] by using KNN with updated embeddings.A dataset called Formpring, proposed by Kelly Reynolds, April Kontostathis, and Lynne Edwards [6], uses machine learning algorithms and oversampling to achieve a recall of 78.5%. This is achieved despite the fact that there is an imbalance in the dataset when it comes to cyberbullying postings. The most recent Google language model, BERST, which produces contextual embeddings for categorization, was used by Jaideep Yadav, Kumar, and Chauhan. With data from form spring, the model produced an F1 score of 0.94; with data from Wikipedia, the score was 0.81. In their study, Sweta Agrawal and Amit Awekar [5] used the same datasets to train Deep Neural Networks.

However, their main emphasis was on using curse words as features for the job. In doing so, they identified platform-specific differences in the lexicon for such models. Building on previous work by Yasin N. Silva, Christopher Rich, and Deborah Hall[6], BullyBlocker is a smartphone app that alerts parents when their kid is the target of cyberbullying on Facebook. The program takes into account warning indicators and susceptibility characteristics to determine the likelihood of cyberbullying occurring.

3. EXISTING SYSTEM

Hsieh [1] obtained a recall of 0.71 and a precision of 0.79 using datasets obtained from four websites by utilizing a technique that included opinion mining, social network analysis, and keyword matching.According to the theory put forward by Patxi Gal'an-Garc'ıa et al. [2], cyberbullies who use social media platforms often maintain a genuine

profile alongside their phony one, in order to gauge how others perceive it. To find these profiles, they suggested using a machine learning method. As part of the identifying procedure, we looked at profiles that are similar to them.

The procedure included picking profiles to analyze, collecting data from tweets, choosing attributes to utilize from profiles, and finally, using ML to identify the tweet's source. A total of 1900 tweets from 19 distinct accounts were analyzed. When it came to author identification, it was 68% accurate. Afterwards, it was used in a Case Study at a Spanish school, where the system was successful in identifying the true owner of a profile among many pupils who were suspected of cyberbullying. A few problems remain with the following approach

As an example, consider a scenario where a trolling account lacks a legitimate account, allowing it to deceive systems or experts that may alter writing styles and behaviors to avoid detection of patterns. More effective algorithms are required for modifying writing styles. In their collaborative detection technique, Mangaonkar et al. [3] suggested using data and outcomes from many interconnected detection nodes, each of which might employ a different or same algorithm. A B-LSTM was proposed by P. Zhou et al. [4]. method that relies on mental focus. According to Banerjee et al. [5]. used KNN together with updated embeddings to achieve a 93% level of accuracy. The downsides It is not possible to construct a vocabulary from every text. Either every word (token) in every document or only the most frequently used ones could make up the vocabulary. Even though it gets its vocabulary characteristics in the same manner as the bag of words model, the Tf-Idf technique is distinct from it.

New Approach In this research, we tackle the subject of cyberbullying detection as a binary classification problem. Specifically, we are looking for two main types of cyberbullying: hate speech on Twitter and personal assaults on Wikipedia. We are trying to determine whether these content types include cyberbullying or not.

In tokenization, we break down raw text into smaller, more manageable pieces called tokens. Tokenizing the sentence "we will do it" into its component parts ('we,' "will,"

"do," and "it") is one example. Word tokenization and phrase tokenization are two different approaches to tokenization. Regex Tokenizer is one of several versions of tokenization that we utilize in this project. The decision-making process for tokens in a regex tokenizer is governed by rules, specifically regular expressions. We choose tokens that match the given regular expression, for example, The regular expression '\w+' is used to extract all the alphanumeric tokens. The term "stemming" refers to the method of breaking a word down into its component parts. For instance, the stem of the triliteral words "eating," "eats," and "eaten" is "eat." It is reasonable to assume that the three terms that stem from the root "eat" mean the same meaning. Porter, Lancaster, Snowball, and Regexp stemmers are the four varieties available from NLTK. One project that makes use of PorterStemmer is this one.

Eliminating superfluous words: In English, superfluous words like "what," "is," "at," and "a" are examples of stop words. You may eliminate these words since they are unnecessary. You may filter out all the tweets using NLTK's collection of English stop words. When training deep learning and machine learning models, it is common practice to delete stop words from text input. This is done because the information that stop words give is useless to the model and may actually improve its performance. The Benefits The Common Bag of Words model takes in a set of input words and makes a word prediction depending on the surrounding text. You may enter a single word or a string of words. The CBOW model averages the input words' contexts, but each word might have two different interpretations selected. in other words, we may anticipate two Apple vectors. The first one is for the fruit, while the second one is for the firm.

4. OUTPUT SCREENS

User:

Userlogin:

Registration:

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

PredictCyberbullying:

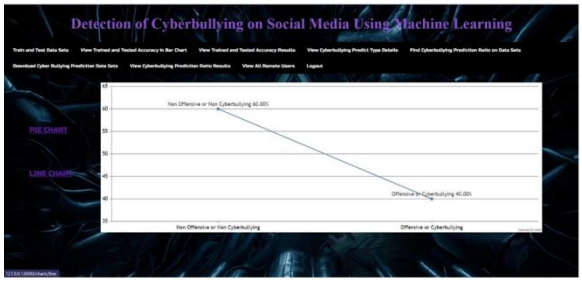
Viewyourprofile:

ServiceProvider:

Adminlogin:

Trainandtestdatasets:

Viewcyberbullyingpredictionratioresults(linechart):



Viewallremoteusers:

5. CONCLUSION

As a result of the serious consequences it causes (e.g., suicide, depression, etc.), it is imperative that cyberbullying be curbed. As a result, identifying

cyberbullying on social media is crucial. More data and improved user information classification means more opportunities for cybercriminals to launch a wide range of assaults. Social media platforms may use cyber bullying detection systems to block users that engage in cyberbullying. In this study, we provide a solution to the problem by outlining an architecture that can identify cyberbullying. Two data sets were covered in our discussion: hate speech data from Twitter and personal assaults data from Wikipedia. Tweets with hate speech were readily identifiable due to the prevalence of profanity, therefore Natural Language Processing approaches using simple Machine Learning algorithms were successful with accuracies of over 90%. For this reason, BOW and TF-IDF models provide superior outcomes compared to Word2Vec models. While all three feature selection approaches worked similarly, it was particularly challenging to utilize the same model to identify personal assaults in comments that lacked a common emotion that could be learnt.Word2Vec models that take feature context into account performed well in both datasets, producing comparable results with less features when paired with Multi Layered Perceptrons.

6.REFERENCES

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L.Wang, and G. M. T.Bermudez, “Towards the detection of cyberbullying based onsocial network mining techniques,” in Proceedings of 4thInternational Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017, 2017, vol. 2018January,doi:10.1109/BESC.2017.256403.

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

[3] A. Mangaonkar, A. and R. “Collaborativedetection of cyberbullying data,” 2015, doi:10.1109/EIT.2015.7293405.

[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection ofcyberbullying on social networks based on bullying features,"2016doi10.1145/2833312.2849567

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

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

[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detectionusing PreTrained doi:10.1109/ICESC48915.2020.9155700.

[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in SocialNetworks Using Deep Learning Based Models; AReproducibility Study," arXiv. 2018.

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

[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards theidentification 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, "Automatedhate speech detection and the problem of offensive language,"2017.

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

[14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis usingdeep 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, "Efficientestimation of word representations in vector space," 2013.