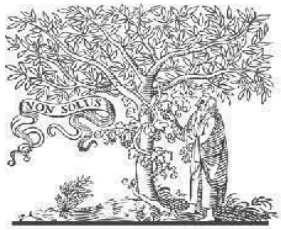
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 3rd October 2024. Link.

https://ijiemr.org/downloads.php?vol=Volume-13&issue=issue10

DOI:10.48047/IJIEMR/V13/ISSUE10/01

Title: **Detection Of Cyber bullying On Social Media Using Machine Learning**

Volume 13, ISSUE 10, Pages: 1-6

Paper Authors
Geetha Prathiba, Jajala. Shirisha, Nalla Spandana, Kolli.Sri Lakshmi

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

DETECTION OF CYBERBULLYING ON SOCIAL MEDIA USING MACHINE LEARNING

[1]Geetha Prathiba, [2]Jajala. Shirisha, [3]Nalla Spandana, [4]Kolli.Sri Lakshmi
[1]Assistant Professor, Department of Information Technology, Malla Reddy Engineering College For Women, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.
[2, 3,4] Student, Department of Information Technology, Malla Reddy Engineering College For Women, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

Abstract_ Cyberbullying is a major problem encountered on internet that affects teenagers and also adults. It has lead to is 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 cyber bullying, hate speech tweets from Twitter and comments based on personal attacks from Wikipedia forums to build a model based on detection of Cyber bullying 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 integralpart of our life. With the evolution of the internet. Social media is trending these days. But as all the other things misusers will pop out sometimes late sometime early but there will be for sure.Now Cyberbullying 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.Cyberbullying 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 cyberbullying??

Cyberbullying is harassment, threatening, embarrassing or targeting someone for the purpose of having fun or even by well-planned means Researches on Cyberbullying Incidents show that 11.4% of 720 young peoples surveyed in the NCT DELHI were victims of cyberbullying 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 Cyberbullying 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 Cyberbullying 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.EXISTING SYSTEM

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 70%.

disadvantages

1.Accuracy is Low.

2. Only using Data Mining algorithms in existing methods .

III.PROPOSED WORK

The proposed device is developing cyberbullying prediction fashions is to use a textual content classification strategy that entails the building of desktop gaining knowledge of classifiers from labeled textual content instances. Another potential is to use a lexicon-based mannequin that includes computing orientation for a record from the semantic orientation of phrases or phrases in the document. Generally, the lexicon in lexicon-based fashions can be developed manually or mechanically by way of the use of seed phrases to make bigger the listing of words. However, cyberbullying prediction the use of the lexicon-based method is uncommon in literature.

The predominant purpose is that the texts on SM web sites are written in an unstructured manner, hence making it tough for the lexicon-based strategy to realize cyberbullying based totally solely on lexicons. However, lexicons are used to extract features, which are frequently utilized as inputs to computer studying algorithms. For example, lexicon based totally approaches, such as the use of a profane- primarily based dictionary to observe the range of profane phrases in a post, are adopted as profane points to laptop getting to know models. The key to high-quality cyber bullying prediction is to have a set of elements that are extracted and engineered

advantages

1.Accuracy is Very high.

2.Better Prediction .

IV. METHODOLOGY

Data Collection:
The first step in the methodology is gathering data from social media platforms. This dataset typically includes a large corpus of user-generated content, such as tweets, posts, comments, and messages, which may contain instances of cyberbullying. Publicly available datasets from platforms like Twitter, Facebook, or Instagram can be used, or proprietary datasets can

be obtained through partnerships with social media companies. The data should be annotated, with each entry labeled as "cyberbullying" or "non-cyberbullying" to facilitate supervised learning. If labeled data is not available, it may need to be manually labeled by experts or using crowd-sourcing techniques.

Data Preprocessing:
Preprocessing is essential to clean and prepare the text data for modeling. This involves several steps: text cleaning to remove unwanted elements such as special characters, punctuation, numbers, and URLs from the text; tokenization, which involves splitting the text into individual words or tokens; stop words removal to eliminate common words that do not contribute much meaning; and stemming and lemmatization, which reduce words to their base or root form (e.g., "running" to "run"). The final step is text vectorization, where the text data is converted into numerical form using techniques such as Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), or word embeddings like Word2Vec, GloVe, or BERT. This allows the machine learning models to process the text data effectively.

Exploratory Data Analysis (EDA):
EDA is performed to understand the distribution and characteristics of the data. This includes analyzing the frequency of cyberbullying-related words, understanding the context in which these words are used, and identifying patterns or trends in the data. Visualization techniques, such as word clouds, bar charts, and histograms, can be employed to gain insights into the most common words or phrases associated with cyberbullying. Additionally, examining the distribution of classes (cyberbullying vs. non-cyberbullying) helps in understanding any class imbalances, which might need to be addressed in the model training phase.

Model Selection:
Various machine learning models are considered for the detection of cyberbullying. Common approaches include Naive Bayes, a probabilistic model well-suited for text classification tasks; Support Vector Machines (SVM), effective for high-dimensional spaces and capturing non-linear patterns; Logistic Regression, a simple and interpretable model for binary classification tasks like cyberbullying detection; and Random Forest, an ensemble method that improves robustness and accuracy by combining multiple decision trees. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) like LSTM (Long Short-Term Memory), are powerful for processing sequential data and

capturing complex patterns in text. Pre-trained language models like BERT, GPT, or RoBERTa, fine-tuned for the specific task of cyberbullying detection, can also be leveraged for enhanced performance.

Model Training and Evaluation:

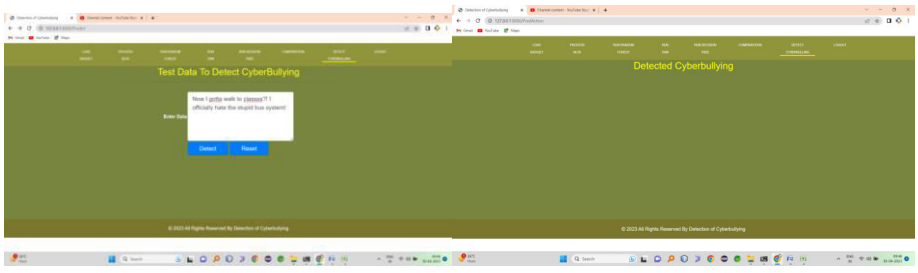
The selected models are trained on the preprocessed text data, with the dataset typically split into training and testing sets using an 80-20 or 70-30 split. Cross-validation techniques, such as k-fold cross-validation, are employed to ensure the model generalizes well to unseen data. The models are evaluated using various performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC (Area Under the Receiver Operating Characteristic Curve). Precision and recall are particularly important in this context, as they indicate how well the model can identify true instances of cyberbullying while minimizing false positives and false negatives. The best-performing model is then selected for further optimization.

Hyperparameter Tuning:

To optimize model performance, hyperparameter tuning is conducted. This involves adjusting the model's parameters, such as the regularization strength in Logistic Regression, the number of trees in Random Forest, or the learning rate in deep learning models. Techniques like Grid Search, Random Search, or Bayesian Optimization are used to identify the best combination of hyper parameters. This step ensures that the model is fine-tuned to deliver the highest possible accuracy and reliability in detecting cyber bullying.

Model Deployment:

Once the model is tuned and validated, it is deployed in a production environment, typically integrated with the social media platform's content moderation system. This enables real-time detection of cyber bullying incidents, where the model analyzes incoming text and flags potentially harmful content for further review or automatic action. The deployment process also includes setting up a monitoring system to track the model’s performance over time, ensuring it continues to operate effectively as new data becomes available.



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 avaibility 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.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 modelsHowever, 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.As seen by changing nature

VI.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 cyber bullying," 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 cyber bullying 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 Cyber bullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.

[6] K. Reynolds, A. Kontosta this, 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,” 201

3