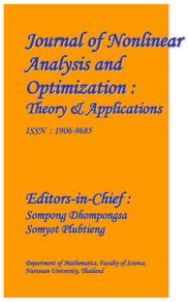
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.15 : 2024
ISSN : **1906-9685**



Paper ID: ICRTEM24_167 **ICRTEM-2024 Conference Paper**

CYBERBULLYING DETECTION USING MACHINE LEARNING

[#1]B. SHRAVYA, ***UG Student,***
[#2]P. SREENIDHI, ***UG Student,***
[#3]K. RAHUL BHARADWAJ, ***UG Student,***
[#4]Dr. G. RAVI KUMAR, ***Associate Professor,***
Department of CSE,
CMR COLLEGE OF ENGINEERING & TECHNOLOGY, HYDERABAD.

ABSTRACT—Hate speech tweets from Twitter and comments based on personal attacks from Wikipedia forums are used in this study 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. Cyber bullying is a major problem encountered on the internet that affects teens and adults. It has led to mis-happenings like suicide and depression. Regulation of content on Social media platforms has become a growing need. For Tweet data the model provides accuracies above 90% and for Wikipedia data it gives accuracies above 80%.

Keywords— *Cyberbullying, Hate speech, Personal attacks, Machine learning, Feature extraction, ML Applications.*

I. INTRODUCTION

More than ever, technology has ingrained itself into our daily lives. With the evolution of the internet. Social media is trending these days. However, just like with everything else, there will undoubtedly be misusers; they may appear early or late. Now Cyberbullying is common these days. Social networking sites are great resources for internal communication. Social networking use has grown throughout time, but generally speaking, people discover unethical and immoral ways to do bad things. This is observed occasionally occurring amongst young adults or teenagers. One of their harmful behaviors is cyberbullying one another. 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 term for using technology to harass, threaten, embarrass, or target someone else. Threats against certain individuals in real life are often the result of this online battle. A few have resorted to suicide. Stopping these kinds of actions right away is essential. Any measures might be implemented to prevent this, such as suspending or terminating an individual's account for a set amount of time if their tweet or message is deemed objectionable. So, what is cyberbullying?? Cyberbullying is harassment, threatening, embarrassing or targeting someone for the purpose of having fun or even by well-planned means.

II. RELATED WORK

In the quest for innovation and efficiency, modern projects frequently rely on existing solutions as fundamental building blocks for development. This approach not only recognizes the expertise and advancements of those who came before us but also nurtures a collaborative ecosystem where ideas can evolve and confront new challenges. In our project, we wholeheartedly embrace this ethos, conscientiously integrating elements from existing solutions to enrich our endeavor. These existing solutions serve as guiding lights, offering insights and frameworks that shape the direction of our project.

A. **Cyber bullying Detection using Pre-Trained BERT Model**. Cyberbullying is spread across various social media platforms. When the bully sends the victim offensive, sensitive, or inflammatory texts or photographs, it is considered harassment and is wrong. It is exceedingly difficult to detect such messages or posts on such vast platforms, and it might occasionally result in false positives. Deep neural network-based models have recently demonstrated notable gains in cyberbullying detection over classical techniques. Additionally, more intricate and sophisticated deep learning architectures are being created, and they are working well for a variety of NLP applications. Recently, BERT, a language learning model created by Google researchers, can create task-specific embeddings for categorization in addition to contextual embeddings. A novel pre-trained BERT model with a single linear neural network layer on top is suggested as a potential method for detecting cyberbullying on social media sites a classifier, which improves over the existing results. Two social media datasets—one of which is tiny in size and the other is comparatively larger—are used to train and assess the model.

B. **Cyber bullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study.** Cyber bullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study. Deep learning-based models have made an appearance in recent studies aimed at detecting instances of cyberbullying. These models make the claim that they can outperform conventional models in terms of detection accuracy and overcome their drawbacks. We examine the results of a recent literature in this regard in this work. By employing the same datasets—Wikipedia, Twitter, and Formspring—as the authors, we were able to replicate their findings and validate them. Then, we extended our research by using the established techniques on a fresh YouTube dataset (around 54k posts from about 4k people), and we looked into how well the models performed on other social media sites. Additionally, we moved the models trained on one platform to another and assessed their performance there as well. Our results demonstrate that the models based on deep learning perform better than the machine learning models that were previously used with the same YouTube dataset. We think that including other information sources and examining the influence of user profile information on social networks can potentially be beneficial for deep learning based models.

C. **Collaborative detection of cyber bullying behaviour in Twitter data.** Unwanted user behaviors on Twitter are growing along with the volume of data it contains. Cyberbullying is one of these unwanted behaviors that might have disastrous results. Therefore, it is imperative to effectively identify instances of cyberbullying through the analysis of tweets, ideally in real time. Common methods for identifying cyberbullying are primarily isolated, which makes them time-consuming. Through the application of collaborative computing techniques, this research enhances the detection task. In this study, many collaboration paradigms are proposed and discussed. According to preliminary findings, the detection process is faster and more accurate than it was with the stand-alone paradigm.

III. METHODS AND EXPERIMENTAL DETAILS

A. Methodology

This project consists of two users

1) **Admin user**: admin can login to application by using username and password as ‘admin’ and ‘admin’. After login admin can view list of users and their count of offensive messages posted by them. If count is more than 2 then Block link will get activated and then admin will block him. Admin can view accuracy and other metrics from trained model.

2) **User:** user can sign up with the application and then can login and then can upload post. While uploading application will detect sentiments and check for offensive words.

Central to our project is the implementation of a blocking model, which enables administrator(admin) to provide warnings to the users who are bullying. Admin plays a pivotal role in issuing the warnings and then blocking the users whose bully count is exceeding 2 within the project, as they are responsible for user’s safety and ensuring accuracy.

The website or system is designed to present the project in a simple and precise manner, aligning with the expectations of students and organizations. Upon receiving user posts, the system will analyze all the comments or messages related to the post and identify the bullies. Thereby facilitating a safe environment.

In summary, our project's methodology revolves around efficient post analysis on social media websites and proper output generation facilitated by interactions. Through streamlined processes and accurate output generation, we aim to enhance the user experience and facilitate seamless access to information.

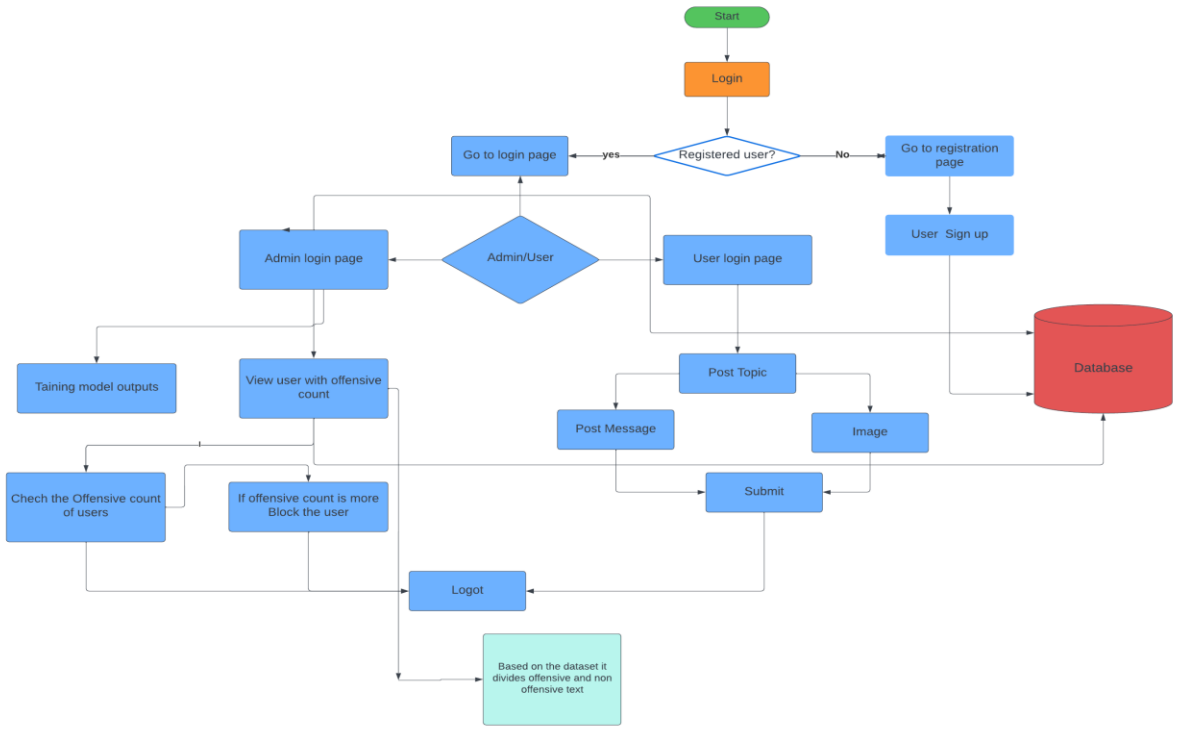


Fig.1.Architecture of the proposed method

B. **Dataset**

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 relies on the use of machine learning and natural language processing to detect cyberbullying in text data.

To determine the optimal strategy, three feature extraction techniques and four classifiers are examined. For Tweet data the model provides accuracies above 90% and for Wikipedia data it gives accuracies above 80%. Using a dataset taken from Twitter, this method was tested; for the positive(cyberbullying) instance, it received an F-score of 0.947.

Twitter datasets contain diverse information collected from the platform, including the text of tweets, metadata such as timestamps and user details, and any attached media like images or videos. These datasets often include information about the users who posted the tweets, including their follower counts and biographical information.

IV. RESULTS AND DISCUSSIONS

The execution of the provided code would yield a detailed evaluation of machine learning models' efficacy in classifying posts as offensive or non-offensive based on sentiment analysis. This assessment encompasses key performance metrics like accuracy, precision, recall, and F1 score, offering valuable insights into the models' classification accuracy. The ensuing discussion would decipher these results, highlighting the strengths and weaknesses of the employed algorithms—AdaBoost, SGD Classifier, and Multinomial Naive Bayes.

Training Approach:
Data Preparation: The application loads preprocessed data (X, Y) from saved files.
Model Training: Three machine learning models (AdaBoost, SGD, Multinomial Naive Bayes) are trained on the data.
Model Persistence: Trained models are saved using pickle for later use.
Model Evaluation: Performance metrics such as accuracy, precision, recall, and F1-score are calculated for each model.
AdaBoost Classifier: A boosting algorithm that combines multiple weak classifiers to create a strong classifier.
SGD Classifier: A linear classifier trained by minimizing a loss function using stochastic gradient descent.
Multinomial Naive Bayes: A probabilistic classifier based on Bayes' theorem with an assumption of independence between features.

Functionality:

User Authentication: Users can register, login, and change passwords. There's also an admin login functionality.
Posting Messages: Users can post messages along with images.
Sentiment Analysis and Cyberbullying Detection: The system analyzes the sentiment of user messages and determines if they contain offensive content using machine learning models.
Admin Panel: An admin panel allows administrators to view offensive posts, block users, and manage user accounts.

Comparison:
Comparing ensemble learning with AdaBoost and other algorithms offers strong performance, ease of interpretation, and scalability, deep learning models can potentially outperform them on certain tasks, particularly when dealing with large and complex datasets. However, deep learning models come with increased complexity and resource requirements. The choice between these approaches depends on factors such as the nature of the data, available computational resources, and the specific requirements of the cyberbullying detection task.

Integration:
By Integrating the provided code into a comprehensive cyberbullying detection system involves orchestrating various components to create a cohesive and functional application. At its core, the system comprises a Django web application augmented with machine learning models for text classification, user authentication, and administrative functionalities.

The Django application serves as the backbone, facilitating user interactions, data management, and model integration. It encompasses views, templates, and models tailored to handle user registration, login, posting messages, and administrative tasks.
Upon user registration, the system captures user details, including username, password, contact information, and profile picture, storing them securely in the database. The profile picture serves as a visual identifier and is stored using Django's Filesystem Storage.
When users post messages, the application preprocesses the text data using techniques like tokenization, stemming, and lemmatization to standardize and clean the input. The preprocessed data is then fed into three machine learning models: AdaBoost, SGD, and Multinomial Naive Bayes.
Each model predicts the sentiment and offensive/non-offensive nature of the message. These predictions guide the system in flagging potentially harmful content and updating the user's offensive count accordingly. Offensive messages trigger actions such as user status updates and administrative notifications.

Below, you'll find a series of images showcasing our innovative web project in action. These visuals provide a glimpse into the user interface, features, and functionality of our web application. Take a moment to explore and discover how our project can prevent bullying in SM platforms. From streamlined interactions to intuitive design, witness firsthand the power and potential of our creation. Dive in and envision the possibilities with our web project.

Fig.2.User Interface

Fig.3.User Signup Screen

Fig.4.User Login Screen

Fig.5.Admin Login Screen

Fig.6.Admin Page

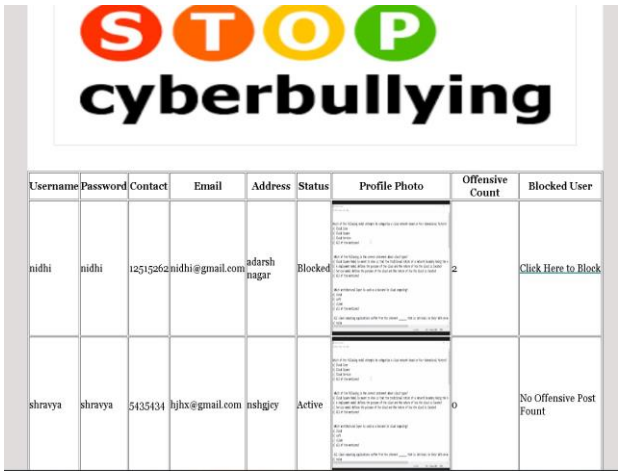


Fig.7.Users with Offensive Count

V. CONCLUSION

In conclusion, the background work and results we have explored offer valuable insights into the potential methodologies for achieving our project's final output. Through the adoption of these methods, we can pave the way for a model that excels in applications while prioritizing precision, engagement, and safety.

Controlling the spread of cyberbullying is necessary since it poses a risk and can result in negative outcomes like depression and suicide. Detecting cyberbullying on social media networks is therefore essential. 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 Social media platforms have the ability to employ cyberbullying detection to prohibit users from engaging in this type of behavior. In order to address the issue, we presented an architecture for cyberbullying detection in this study. We discussed the architecture for two types of data: Personal attacks on Wikipedia and hate speech data on Twitter. 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.

With the completion of this project, we envision a cyberbullying detection model that not only meets but exceeds the expectations of users in engineering contexts. By leveraging the methodologies discussed and implementing

them effectively, we are poised to deliver a solution that empowers users, enhances productivity, and fosters innovation in engineering domains.

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 SOCIOCULTURAL 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.