Social media Based cyber bullying detection

Jeevana Priya Chintala, Dr. A. Radhika
Department of Computer Science and Engineering, SRK Institute of Technology, Vijayawada, Andhra Pradesh, INDIA
jeevanapriya710@gmaijl.com
Professor, Department of Computer Science and Engineering, SRK Institute of Technology, Vijayawada, Andhra Pradesh, INDIA

ABSTRACT
Increasing internet use and facilitating access to online communities such as social media have led to the emergence of cybercrime. Cyber bullying is a pervasive issue on the internet, affecting both teenagers and adults, and leading to severe consequences such as suicide and depression. The need for stricter content regulation on social media platforms has become increasingly apparent. This study addresses the problem of cyber bullying by utilizing data from two different forms: hate speech tweets from Twitter and comments containing personal attacks from Wikipedia forums. The research focuses on building a robust cyber bullying detection model using Natural Language Processing and machine learning. Three distinct methods for feature extraction and four classifiers are thoroughly examined to identify the most effective approach. This proposed system aims to provide a proactive solution for the detection and prevention of cyber bullying, thereby promoting a safer online environment
Keywords: Machine Learning, Cyberbullying, Social Media, Twitter**.**

INTRODUCTION
Cyber bullying is bullying online. Most of the time, if not all the time over social media. Cyber bullying is often not physical, which means that the people being cyber bullied feel mental pain instead of physical. Cyber bullying started when social media was made, and there's been more and more cases since. Cyber bullying expresses the acts of the slanders, gossips, threaten harassment, insult, abashing and excluding someone on the digital world. It is a new generation bullying. Cyber bullying has influence on people more psychological. Seventy percent of people are exposed to cyber bullying every year .It is usually being done via fake accounts. Some people who are exposed to cyber bullying are thinking of suicide. Authorities and parents have to talk to children about cyber bullying .They should listen to them without judging. Adults should be role model about respecting the others. Children have to be taught to fight with cyber bullying. People who are suffering from cyber bulling should be encouraged to talk about what they live. We mustn't be a cyber-bully. Cyberbullying on social media platforms is a pervasive issue with limited and often ineffective content moderation Mechanisms. The main drawbacks of the existing system are Limited Detection Capabilities that is the existing system relies on basic and outdated methods, resulting in suboptimal cyberbullying detection capabilities, often failing to identify nuanced forms of online harassment. Due to the lack of real-time monitoring and preventive measures, the current system provides delayed responses, allowing cyberbullying incidents to escalate before any intervention occurs, so to address the shortcomings of the existing system by employing advanced Natural Language Processing and machine learning techniques to detect and prevent cyberbullying more effectively. To perform this the present existing system follows different plan of actions such as Gathering extensive datasets from various social media platforms, including Twitter and Wikipedia, containing examples of cyberbullying. Data Preprocessing to implement data cleaning, text normalization, and feature extraction techniques to prepare the data for analysis. From those preprocessed data different methods are explored for feature extraction that includes TF-IDF, word embeddings, and sentiment analysis. Model Selection is done by evaluating multiple machine learning algorithms and

classifiers to identify the most effective ones for cyberbullying detection. Model Training is done by training the selected models on the preprocessed data and fine-tune hyperparameters for optimal performance. Later evaluation is done using evaluation Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess model performance. Real-time Detection is done by implementing a real-time cyberbullying detection system for social media platforms. Develop mechanisms for reporting and preventing cyberbullying incidents.

LITERATURE SURVEY

Al-Ajlan(2018)[1]:The prevalence of cyberbullying on social media processes the development of robust predictive models to detect and prevent harmful behavior online. Machine learning and deep learning-based approaches show great promise in addressing this challenge, offering scalable and efficient solutions for cyberbullying detection. However, the effectiveness of these models hinges on addressing key challenges such as data quality, model interpretability, and ethical considerations. Moving forward, interdisciplinary efforts involving researchers, policymakers, and social media platforms are essential to combatting cyberbullying and fostering a safer online environment for all users.

Mahlangu, T., Tu, C(2019) [2]: Cyberbullying remains a significant challenge in the digital age, necessitating proactive measures to detect and prevent harmful behavior on social media platforms. This systematic literature review provides valuable insights into the state-of-the-art methods, benchmark datasets, and challenges in cyberbullying detection research. By addressing these challenges and leveraging advanced computational techniques, researchers can continue to advance the development of effective cyberbullying detection systems, ultimately fostering safer and more inclusive online communities

.

Prosun, P.R.K (2021)[3] : This paper offers a detailed examination of cyberbullying, legal implications, and prevention strategies, particularly within the context of Indian college settings. Despite limitations like limited empirical data and potential subjectivity in expert opinions, it underscores the urgency for tailored regulations to combat cyberbullying effectively. Collaboration among policymakers, legal experts, and stakeholders is crucial for implementing preventive measures and fostering safer online environments in educational institutions across India.

Dewani. A (2021)[4] : This study sheds light on the overlooked issue of cyberbullying faced by teachers on social media platforms within educational settings in Nepal. Through qualitative methods, it provides valuable insights into the types of cyberbullying experienced by teachers and their coping strategies. The findings underscore the need for robust policies and strict cyber laws to protect teachers' cybersecurity and ensure their well-being. While the study's limited sample size and potential subjectivity in data collection are noted, its practical implications for policymakers and educational institutions highlight the urgency of addressing cyberbullying and safeguarding teachers in the digital age.

Shen, W. (2012)[5] : This scoping review highlights the prevalence and impact of cyberbullying on children and adolescents' mental health, particularly in the context of social media. While it identifies a consistent association between cyberbullying and depression, further research is needed to understand its effects on other mental health conditions. The findings underscore the importance of developing effective prevention and management strategies, considering the passive responses often observed among recipients. However, the review's focus on studies primarily conducted in the United States may limit its global generalizability. Nonetheless, it provides valuable insights for informing future research and

interventions aimed at addressing cyberbullying and promoting the well-being of children and young people.

Rahman, M (2023)[6]: This cross-sectional study provides insights into the indirect effects of traditional and cyber peer victimization on the relationship between ADHD symptoms and sleep disturbance/impairment among elementary school students. The findings highlight the potential mechanisms linking these variables and suggest implications for interventions aimed at reducing sleep problems in children with ADHD symptoms. However, limitations such as reliance on self-report measures, the cross-sectional design, and the sample's limited generalizability should be considered when interpreting the findings. Nonetheless, the study contributes to our understanding of the interplay between ADHD symptoms, peer victimization, and sleep issues, offering practical implications for supporting children's psychosocial well-being in real-world settings.

Mahat, M (2020)[7] : This paper provides an overview of methods and challenges in cyberbullying prevention, highlighting the importance of addressing this pervasive issue in the digital age. While various techniques have been proposed to combat cyberbullying, challenges such as anonymity, psychological impact, and the need for coordinated efforts persist. By identifying areas for future research and continued innovation, the paper aims to contribute to the development of effective interventions and policies to promote online safety and prevent cyberbullying.

Joshi, A.(2021)[8] : This human-centered systematic literature review sheds light on the past decade of research on automated cyberbullying detection. While the findings indicate progress in technical innovation, there remains a significant gap in incorporating human-centeredness across various aspects of detection systems. Moving forward, it is crucial for future research to prioritize the integration of human perspectives, beliefs, and needs to develop detection systems that are more practical, useful, and attuned to the diverse contexts and sensitivities of stakeholders involved in cyberbullying incidents.

PROJECT AIM

The main aim of the detecting the cyberbullying model will help to improve manual monitoring for cyberbullying on social networks. In this project we fetch the tweets from twitter accounts and preprocess the twits and images and applying generated model will detect the cyberbullying or not. The objectives of the systems development and event management are: Collect the data set of bullying words and preprocess it and apply natural language processing and then machine learning algorithms Generate different machine learning algorithm model. Fetch the tweets from twitter account and preprocess it. Apply generated model on the fetched tweets and get final output cyberbullying or not.

PROJECT SCOPE

Cyberbullying is the use of electronic communication to bully a person by sending harmful messages using social media, instant messaging or through digital messages. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying. So overcome these issues detecting the cyberbullying is very important in now a days which will help to stop cyberbullying on social media networks.

search and find the the dataset and download it for train the model. After downloading first we will pre process the data and then transferred to Tf Idf. It then trains the dataset using Naive Bayes, Support Vector Machine (SVM), and DNN algorithms and creates models separately. Next, we will develop a web application using the FLASK framework. It imports live tweets from Twitter, then applies the generated model to the imported tweets and checks whether text or images are cyberbullying. For this purpose, we use python as backend, Mysql as database and html, css, javascript etc as frontend.

METHODOLOGY

Multinomial Naive Bayes Classifier*:*

Multinomial Naïve Bayes classifier is used to classify the type of bullying. With training, the algorithm classifies cyber bullying as-Shaming, Sexual harassment and Racism. Experimental results show that the accuracy of the classifier for considered data set multinational Naïve Bayes classifier is used to classify the type of bullying.

K Nearest Neighbor (KNN)

In this method reads the tweets and then classifies the texts relating to cyber bullied and blocks the users. The study uses a k-NN classifier integrated with Deep Learning and show how effective the model is over large text datasets than other methods.

Random Forest Classifier:

The wrong prediction is caused by the inconsistency in which some rules that determine whether a group of tweets are considered to be cyberbullying the most are justified as non-cyberbullying in some cases as well as the other way around.
This gives high accuracy

XGBoost Classifier:

In finding a negative comment from the messages it receives by a user. The algorithm first gives the message a value and then based on our pre trained data, it decides if the comment is harsh enough to be transformed or not.This gives modarate accuracy

ADABoost Classifier:

Combines the predictions of multiple weak learners to produce a strong classifier with improved accuracy.A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on FormSpring in Textual Modality.

PROBLEM STATEMENT

Social networks Networks give us great opportunities to communicate, and also increase the vulnerability of young people to threatening situations on the Internet. Cyberbullying on social media is a global phenomenon due to its large number of active users. The trend shows that social network cyberbullying is increasing rapidly day by day. Recent research shows that cyberbullying is a growing problem among young people. Successful prevention depends on the proper detection of potentially harmful messages, and the information overload of the Internet requires intelligent systems to automatically detect potential hazards. Therefore, in this project, we will focus on creating a model to automatically detect cyberbullying in social media text by simulating messages created by social media bullying.

OUTPUT OF CYBERBULLYING DETECTION SYSTEM

In this screenshot we have shown the detection plane shows an example of bullying content that is detected as bullying in the result," This is a type of bullying content", and if the text is not the content of cyberbullying there will be a output is "This is not a type of bullying content".
As the framework is done in the Django show that "USER NAME,ENTER PASSWORD,RE-ENTER PASSWORD, ENTER EMAIL, ENTER PHONE NUMBER " for register option and

if the user have the account then they can login with already created account with " LOGIN" option as shown in Fig. 1.

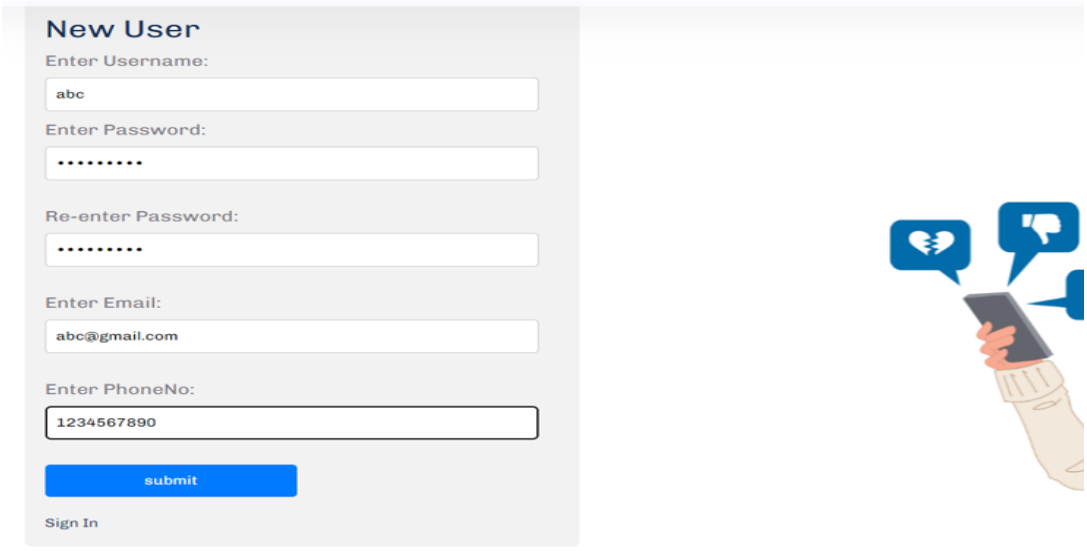


Figure 1: login page

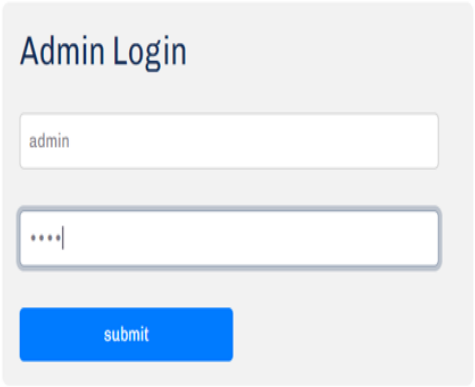
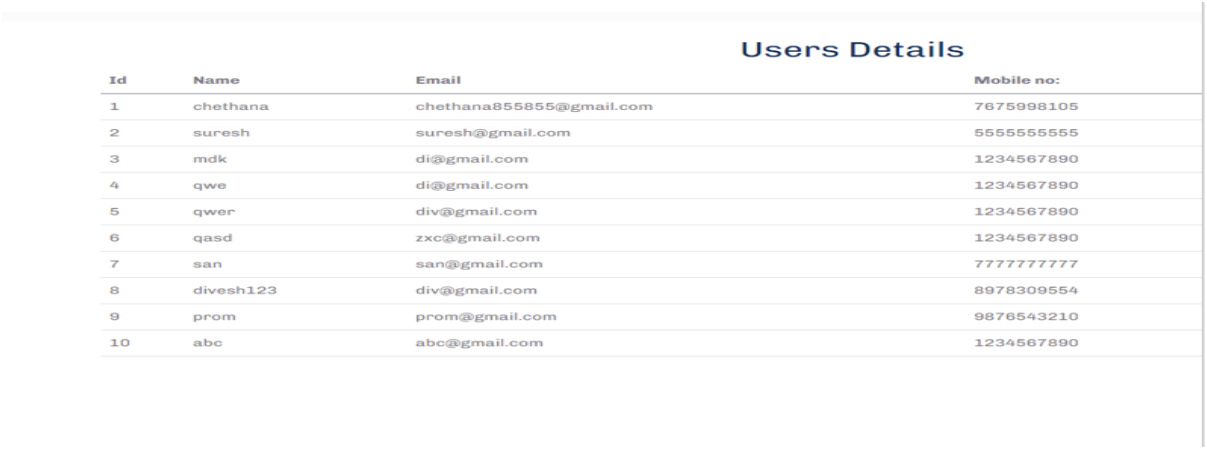


Figure2: admin login

Step 2 which is the user admin login page who accepts the new user and activates that profile through that option lies beside the user sign in that transfers to the next page which is opens in the next tab who activates that profile and gives them permission to use the prompt. This can be seen in fig.2



Users Details

Id	Name	Email	Mobile no:
1	chethana	chethana855855@gmail.com	7675998105
2	suresh	suresh@gmail.com	5555555555
3	mdk	di@gmail.com	1234567890
4	qwe	di@gmail.com	1234567890
5	qwer	div@gmail.com	1234567890
6	qasd	zxc@gmail.com	1234567890
7	san	san@gmail.com	7777777777
8	divesh123	div@gmail.com	8978309554
9	prom	prom@gmail.com	9876543210
10	abc	abc@gmail.com	1234567890

Figure 3: User details

That specific user gets the activation of that profile and ready to use environment for that specific person that too they are eligible to use them when they have valid credentials which are EMAIL and PASSWORD respectively which is seen in fig.3

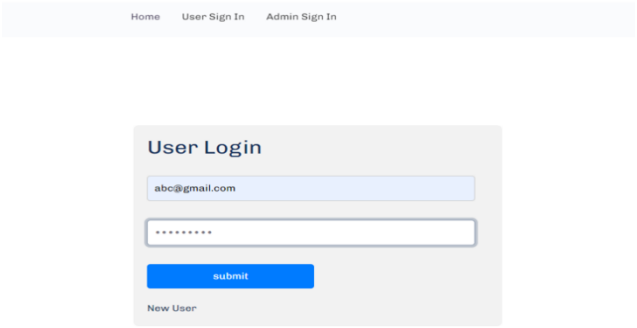


Figure 4: User details

Then that user gets sign in using his mail and password to that log in page visible in the fig.4

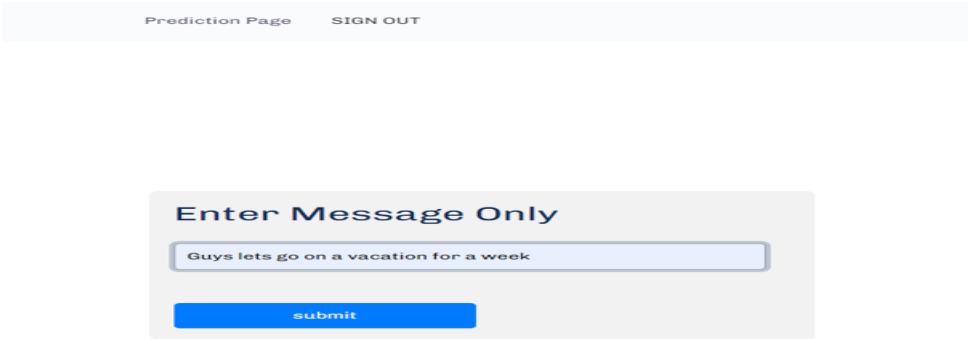


Figure5: input message

Shows a message bar for text insertion and then it shows that specific version presetting the status of that category which is seen as model prediction output and showcases weather it could be a reported or not. Seen in fig.

Figure 6: output

Shows the message is bullyed or not under which type it is divided as we are divided it in to that specific verified version. Seen in fig.6

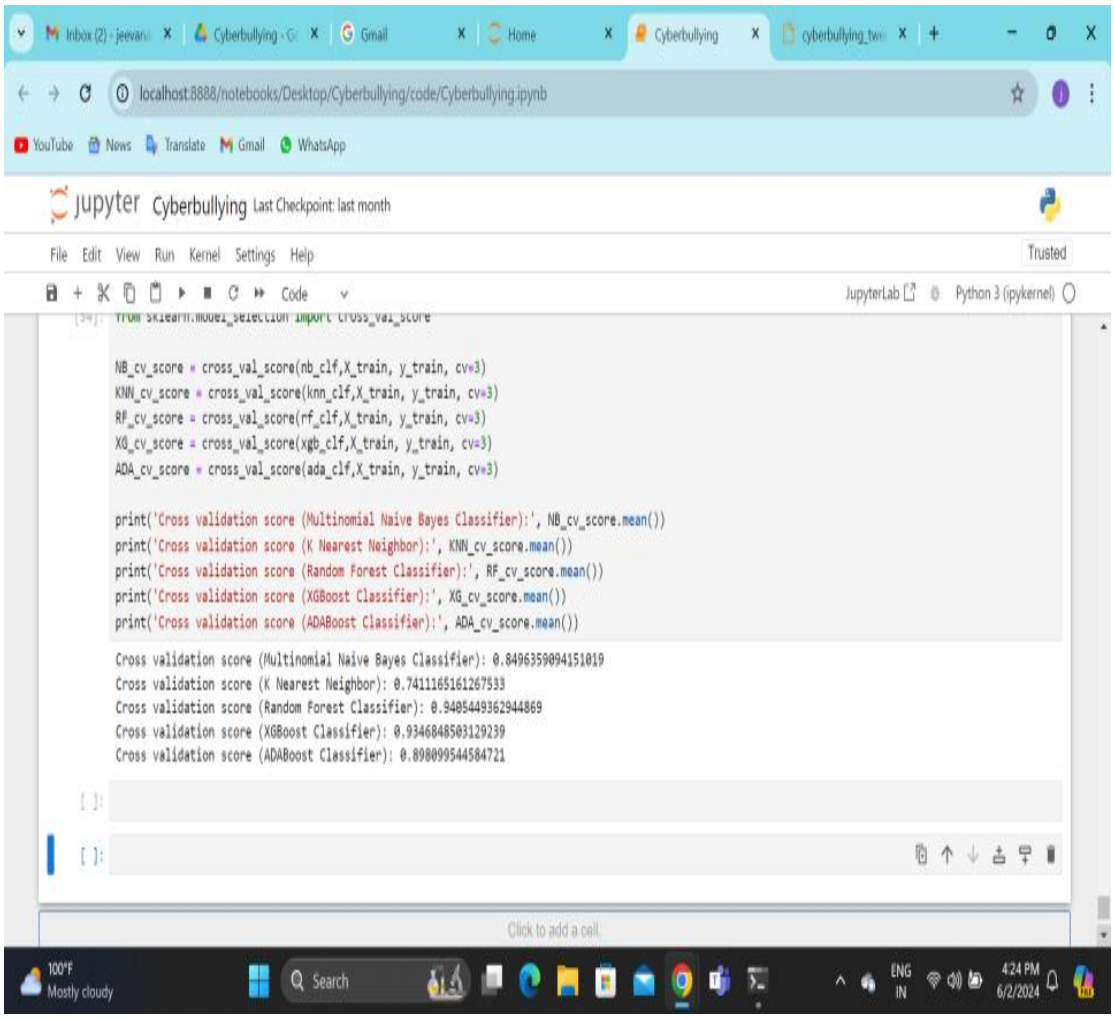


Figure 7:highest accuracy

The cyberbullying detection is done using multinomial naïve bayes classifier,K nearest Neighnor,Random Forest Classifier,XGBoost Classifier,ADABoost Classifier algorithms.Accuracy metrices are found and of those the algorithms which gives the highest accuracy is XGboost and random forest when compared with the taken list of algorithms

ADVANTAGES

Cyber bullying detection process is automatic and time taken for detection is less.It works on live environment. The latest machine learning models are used for training models that are accurate.

APPLICATIONS

This software application would be capable of accurately classifying Twitter message negative or positive with respect to some commonly used terms. Mainly focused on gender bullying by using four categories with different polarity.

CONCLUSION

We proposed a semi supervised approach in detecting cyberbullying based on the five features that can be used to define a cyberbullying post or message using the BERT model. While considering just one of the features which was sentimental features the BERT model achieved 91.90% accuracy when trained over dual cycles which outperformed the traditional machine learning models. The BERT model can achieve more accurate results if provided with a large data set. We can try to achieve even better results in the cyberbullying detection process if we consider all the features that we have proposed in this research paper. Based on all the features an application can be created to detect the bullying traces and thus help in detecting and reporting such posts. A combination of other models on top of the BERT model may be used in the future to generate a heart condition model for a specific NLP cyberbullying detection task.

REFERENCES

[1.]Al-Ajlan, M.A., Ykhlef, M.: Optimized twitter cyberbullying detection based on deep learning. In: 2018 21st Saudi Computer Society National Computer Conference (NCC), pp. 1–5 (2018). IEEE

[2.]Mahlangu, T., Tu, C.: Deep learning cyberbullying detection using stacked embbedings approach. In: 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 45–49 (2019). IEEE

[3.]Alam, K.S., Bhowmik, S., Prosun, P.R.K.: Cyberbullying detection: an ensemble based machine learning apcvproach. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 710–715 (2021). IEEE

[4.]Dewani A, Memon MA, Bhatti S. Cyberbullying detection: a preprocessing techniques & deep learning architecture for roman urdu data. J Big Data. 2021;8(1):1–20.[5.]Luo, Y., Zhang, X., Hua, J., Shen, W.: Multi-featured cyberbullying detection based on deep learning. In: 2022 16th International Conference on Computer Science & Education (ICCSE), pp. 746–751 (2021). IEEE

[5.]Yadav, J., Kumar, D., Chauhan, D.: Cyberbullying detection using pre-trained bert model. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1096–1100 (2020). IEEE

[6.]Ahmed, M.T., Rahman, M., Nur, S., Islam, A., Das, D.: Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study. In: 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1–10 (2021). IEEE

[7.]Mahat, M.: Detecting cyberbullying across multiple social media platforms using deep learning. In: 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 299–301 (2021). IEEE

[8.] Jain, N., Hegde, A., Jain, A., Joshi, A., Madake, J.: Pseudo-conventional approach for cyberbullying and hate-speech detection. In: 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), pp. 1–8 (2021). IEEE.