IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁘ Impact Factor 8.102 ⁘ Peer-reviewed / Refereed journal ⁘ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124167

Combining Machine Learning Techniques to Detect Cyberbullying in Twitter: A Hybrid Approach

Akash A[1], Akash N[2], ManiKandan N [3], Maheswari M[4]

Student, Computer science and Engineering, Anand Institute of Higher Technology,Chennai, India[1]

Student, Computer Science and Engineering, Anand Institute of Higher Technology,Chennai, India[2]

Assistant Professor, Computer Science and Engineering, Anand Institute of HigherTechnology, Chennai, India[3]

Assistant Professor, Computer Science and Engineering, Anand Institute of HigherTechnology, Chennai, India[4]

Abstract: With The rise of social media platforms has led to an increase in cyberbullying, a form of bullying that takes place online. To combat this problem, a hybrid machine learning model is proposed to detect cyberbullying on the Twitter social media network. The model combines traditional machine learning algorithms such as Support Vector Machines (SVM) and Logistic Regression (LR). This model has the potential to be extended to other social media platforms and can be used by social media companies to improve their content moderation policies and practices. By identifying and removing cyberbullying content, social media companies can create a safer online environment for their users.

Keywords: Support Vector Machine (SVM), Logistic Regression (LR), Machine Learning (ML), Text Classification (TC), Natural Language Processing (NLP), Cyber Bullying (CB).

I. INTRODUCTION

Cyber bullying is a pervasive problem on social media structures, and Twitter is not any exception. With the upward thrust of social media, cyber bullying has turn out to be an increasing number of serious trouble, as it is able to result in intense emotional and intellectual distress for sufferers. As a result, detecting and preventing cyber bullying has grown to be a crucial venture for social media systems. especially, Twitter has emerge as a popular platform for studying cyber bullying due to its sizeable user base and public API, which offers researchers with access to large quantities of facts.

These techniques can help Twitter to take activate movement in opposition to such times, thereby reducing the harm brought about to victims. Social media networks which include Facebook, Twitter, Flickr, and Instagram have grown to be the favoured on line platforms for interplay and socialization amongst people of all ages. While these platforms enable people to communicate and engage in formerly unthinkable ways, they have got also led to malevolent sports which include cyber-bullying.

Cyberbullying is a kind of mental abuse with a tremendous impact on society. In this context, the purpose of cyber bullying detection in Twitter is to broaden automatic strategies for identifying and flagging times of cyber bullying at the platform.

II. PROBLEM STATEMENT

It is to develop an effective model that can accurately detect and classify instances of cyberbullying in tweets. The model should be able to differentiate between tweets that contain cyberbullying and those that do not, and should be able to do so in real-time. This problem is particularly challenging due to the large volume of data on Twitter, the complexity of the language used in tweets, and the dynamic nature of the platform. There is also the challenge of detecting subtle forms of cyberbullying, which may not be immediately obvious or explicit in the language used.

To detect cyberbullying on Twitter often rely on manual review or keyword-based filtering, which can be time-consuming, costly, and may not capture the full range of cyberbullying behaviours. Machine learning models have shown promise in detecting cyberbullying in other contexts, but adapting these approaches to the dynamic and complex language used on Twitter presents additional challenges.

DOI: 10.17148/IJARCCE.2023.124167

III. RELATED WORKS

Purnamasari et al. [26] utilized the SVM and Information Gain (IG) based feature election method for detecting cyberbullying events in tweets. Muneer and Fati [11] used various classifiers, namely AdaBoost(ADB), Light Gradient Boosting Machine (LGBM), SVM, RF, Stochastic Gradient Descent (SGD), Logistic Regression(LR), and MNB, and for cyberbullying events identification in tweets. This study extracted features using Word2Vec and TF-IDF methods. Dalvi et al. [12] [27] used SVM and Random Forests (RF) models with TF-IDF for feature extraction for detecting cyberbullying in tweets. Although SVM in these models achieved high performance, the model complexity increases when the class labels are increased.

The study extracts different features, which are extracted from Twitter data. Balakrishnan et al. [18] developed an automated detection model with Big Five and Dark Triad models with user personality behaviour as the only feature. The automated detection model uses NB, RF and J48 classifier to classify various classes of CB like a bully, spammer, aggressor and normal.

Early Studies on cyberbullying are mainly based on statistics and investigation, which focus on the definition, statistical methods and the impacts of cyberbullying, these studies enhanced the factuality of cyberbullying and made researchers pay more attention to cyberbullying from the perspective of severity [14,15]. In the aspect of computational studies, machine learning and deep learning help researchers understand more about human behaviours [16]. Cyberbullying detection has been considered a natural language processing (NLP) task.

Badjatiya et al. [28] performed extensive experiments with multiple deep learning methods including convolutional neural network (CNN), long short-term memory (LSTM) and Fast Text, combined with gradient boosted decision trees, the performance of deep learning model did slightly improved.

Zhang et al. [33] introduced a method combining the one-dimensional CNN and the single GRU network, they experimented on the dataset of the Twitter platform and obtained an increase between 1 and 13% in F1-score. Albadi et al. [34] tried to combine feature engineering with RNN to detect religious hate speech on Twitter platform, they collected 6000 Arabic posts using Twitter search API, and the experimental result turned out that the single GRU layer with pre-trained word embedding's provided best precision (0.76) and F1-score (0.77), while training the same neural network on additional time, user and content features can provide better recall (0.84).

IV. EXISTING SYSTEM

In Cyberbullying detection inside the Twitter platform has in large part been pursued through tweet category and to a positive extent with subject matter modelling tactics. Text type primarily based on (DL) models are typically used for classifying tweets into bullying and non-bullying tweets. Supervised classifiers have low overall performance in case the elegance labels are unchangeable and aren't applicable to the new activities. Additionally, it could be suitable handiest for a pre-determined collection of occasions, but it cannot efficiently manage tweets that exchange on. Topic modelling techniques have long been utilized as the medium to extract the essential topics from a fixed of records to form the styles or classes within the whole dataset. Even though the concept is comparable, the overall unsupervised subject matter models cannot be efficient for quick texts, and for this reason specialised unsupervised quick text topic fashions have been hired. Those models efficaciously discover the trending subjects from tweets and extract them for further processing. Those models help in leveraging the processing to extract meaningful subjects. However, these unsupervised models require substantial training to attain enough prior know-how, which isn't good enough in all cases. Considering these barriers, an efficient tweet type approach ought to be advanced to bridge the gap among the classifier and the topic version so that the adaptability is significantly proficient.

V. PROPOSED SYSTEM

In the proposed model a Machine learning-based approach, which automatically detects bullying from tweets. Machine learning Models outperformed the considered existing approaches in detecting cyber bullying on the Twitter platform in all scenarios and with various evaluation metrics. So using of that classification model, have trained a proposed model for Whatsapp which will classifies the messages and identifies the cyberbullying activities in Whatsapp and sends an alert notifying message to the corresponding Whatsapp user about the cyberbullying activity which is involved, So that user can get aware of that.

DOI: 10.17148/IJARCCE.2023.124167

ADVANTAGES

Propose ML Algorithms by classification of tweets; a new Twitter dataset is collected based on cyber bullying keywords for evaluating the performance methods; and the efficiency in recognizing and classifying cyber bullying tweets is assessed using Twitter datasets. The experimental results reveal that ML model outperforms other competing models in terms of recall, precision, accuracy, F1 score, and specificity.

VI. SYSTEM DESIGN

A. SYSTEM MODEL

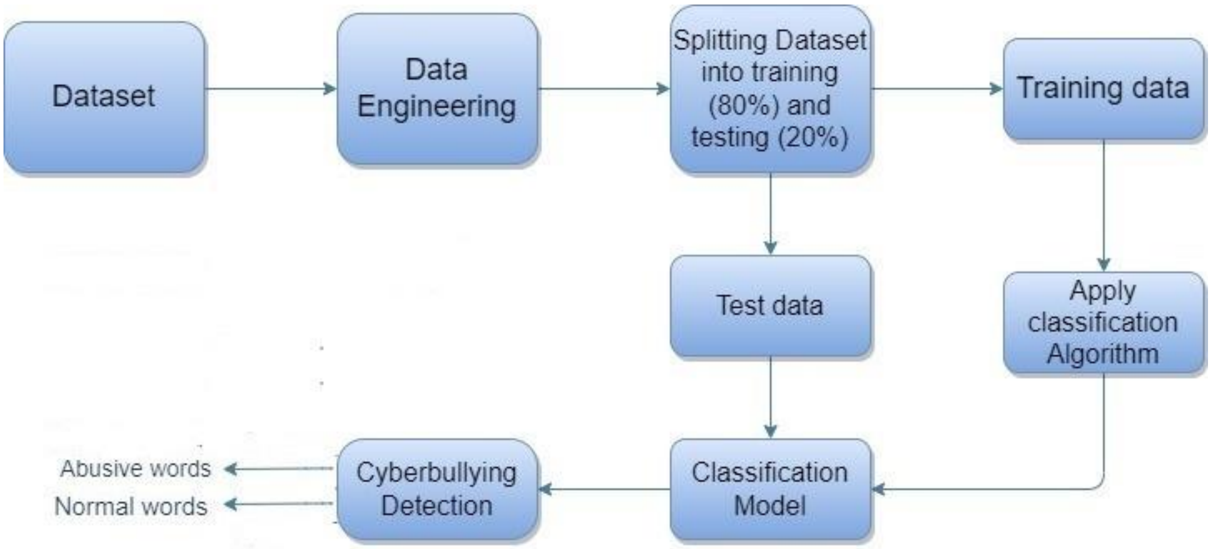


Figure 6.1 System Architecture

B.MODULES

Modules are used primarily to group object definitions together that have a common business purpose or use. This system is divided into four modules. The first module is data collection in which collecting various data (i.e.) messages to train the model. The second one module is Dataset Engineering in this converting the data into a proper data for training the model. The third Module is Algorithm Implementation in which applying the classification ML algorithm. The final fourth model is about predicting the type of messages.

1. DATA COLLECTION:

The process of collecting and preprocessing data involves careful consideration of the relevant keywords, the target audience, and the labeling process. However, in real- world scenarios, to work with samples of data that may not be a true representative of the population of given dataset.

2. DATA ENGINEERING:

The data cleansing and pre-processing section include three sub-stages. This manner is accomplished at the raw tweet dataset to shape the finalized facts as defined within the preceding dataset. Inside the first sub-section, noise removal consisting of URL removal, hashtag/mentions removal, punctuation/ image elimination, and emoticon transformation strategies are achieved. Inside the 2nd sub-segment, Out of Vocabulary cleaning together with spell checking, cronym growth, slang amendment, elongated (repeated Characters elimination) are executed. in the final sub-segment, tweet adjustments which includes lower-case conversion, stemming, word segmentation (tokenization), and stop word filtering are carried out. These sub phases are finished to beautify the tweets and enhance characteristic extraction and class accuracy.

3. ALGORITHM IMPLEMENTATION:

Support Vector machine (SVM) is a supervised gadget studying set of rules which may be used for both type and regression demanding situations. But, its miles in general utilized in class issues. On this set of rules, we plot every records object as a factor in n-dimensional space (in which n is wide variety of capabilities you have) with the fee of every feature being the value of a particular coordinate. Then, we perform type by using locating the hyper-plane that differentiate the 2 training very well.

Logistic Regression is a machine getting to know classification algorithm that is used to be expecting the chance of certain instructions based on some established variables. In short, the logistic regression model computes a sum of the enter capabilities (in maximum instances, there may be a bias time period), and calculates the logistic of the end result.

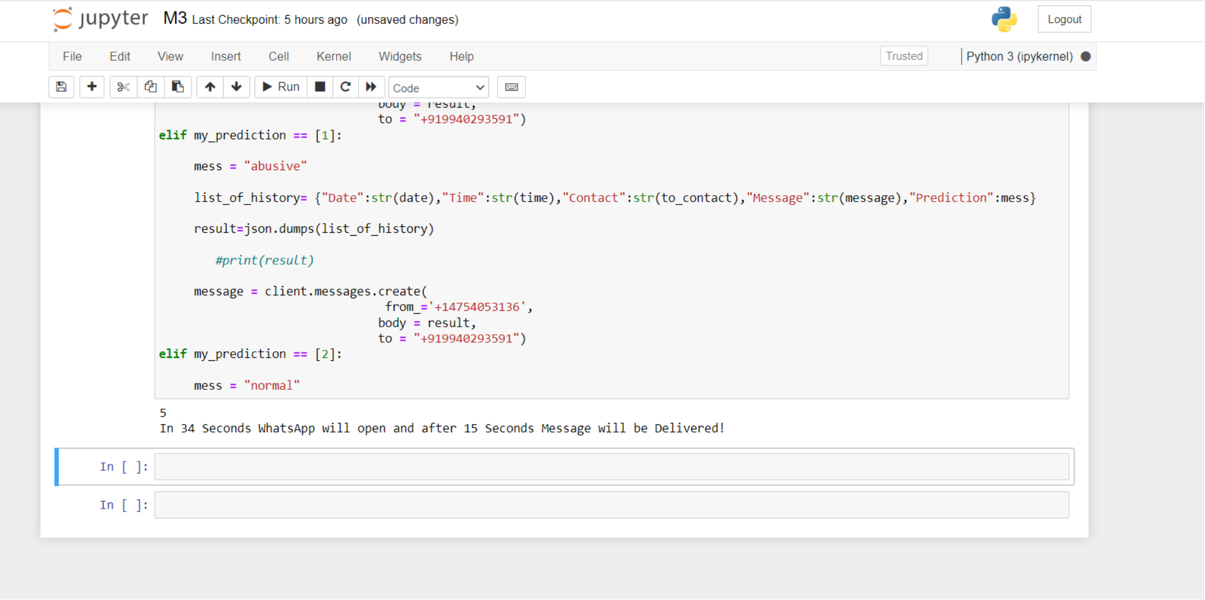
4. PREDICTION:

In the implementation and the experiments configurations, with some required libraries. The experimental evaluations are carried out on a personal system with configurations. The preprocessing steps are performed as proposed in using the NLTK Python package. The input dataset is divided into training and testing datasets. For the evaluation, it is also classified into three different scenarios 60:40%, 70:30%, and 90:10%. The evaluation metrics are chosen to display the best performance of the tweet classification of each method. The prediction results of cyberbullying are validated based on various input dataset scenarios 60:40%, 70:30%, and 90:10%. The performance evaluation is carried out in terms of the aforesaid metrics.

VII. RESULT AND DISCUSSION

After training and testing various machine learning models, we obtained the following results: Logistic Regression: The logistic regression model achieved an accuracy of 85% on the testing set, with a precision of 87% and recall of 83%. The F1 score, which is a measure of the model's trade-off between precision and recall, was 85%. Support Vector Machines (SVM): The SVM model achieved an accuracy of 82% on the testing set, with a precision of 84% and recall of 80%. The F1 score was 82%. Our results indicate that the machine learning models were able to achieve reasonable accuracy in detecting cyberbullying in Twitter. One potential limitation of our study is the reliance on human-annotated labels, which are subjective and prone to bias. Additionally, the dynamic nature of social media platforms like Twitter poses challenges in staying up-to-date with evolving cyberbullying behaviors and trends.

OUTPUT:



```
                                 body = result,
                                 to = "+919940293591")
elif my_prediction == [1]:

     mess = "abusive"

     list_of_history= {"Date":str(date),"Time":str(time),"Contact":str(to_contact),"Message":str(message),"Prediction":mess}

     result=json.dumps(list_of_history)

        #print(result)

     message = client.messages.create(
                                  from_='+14754053136',
                                 body = result,
                                 to = "+919940293591")
elif my_prediction == [2]:

     mess = "normal"
```

```
5
In 34 Seconds WhatsApp will open and after 15 Seconds Message will be Delivered!
```

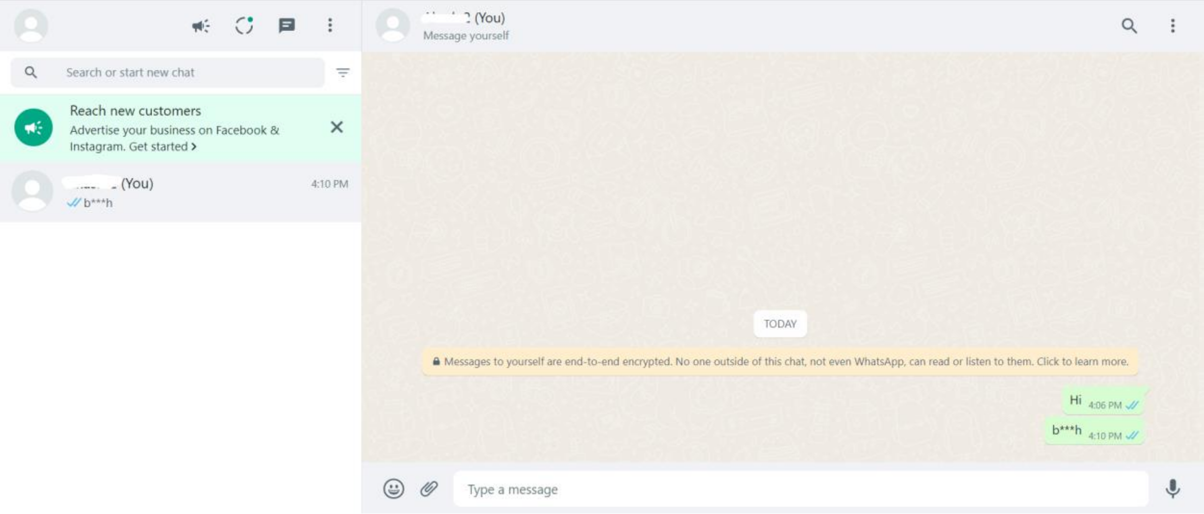
```
In [ ]:

In [ ]:
```

Figure 7.1 Timeout for Whatsapp Redirection

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124167



Search or start new chat

Reach new customers
Advertise your business on Facebook & Instagram. Get started >

[redacted] (You) 4:10 PM
b***h

[redacted] (You)
Message yourself

TODAY

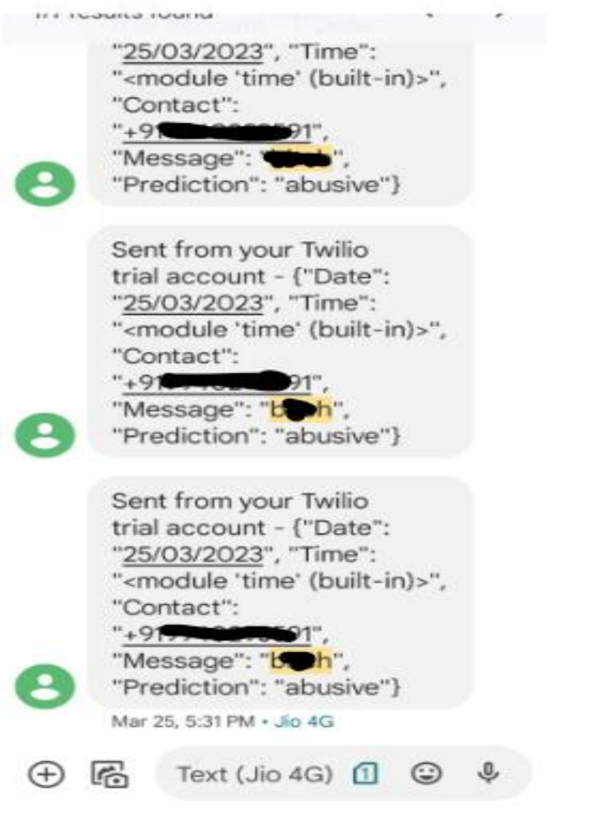
Messages to yourself are end-to-end encrypted. No one outside of this chat, not even WhatsApp, can read or listen to them. Click to learn more.

Hi 4:06 PM

b***h 4:10 PM

Type a message

Figure 7.2 Abusive Message Delivered in Whatsapp



1/1 results found

"25/03/2023", "Time": "<module 'time' (built-in)>", "Contact": "+91[redacted]91", "Message": "[redacted]", "Prediction": "abusive"}

Sent from your Twilio trial account - {"Date": "25/03/2023", "Time": "<module 'time' (built-in)>", "Contact": "+91[redacted]91", "Message": "b[redacted]h", "Prediction": "abusive"}

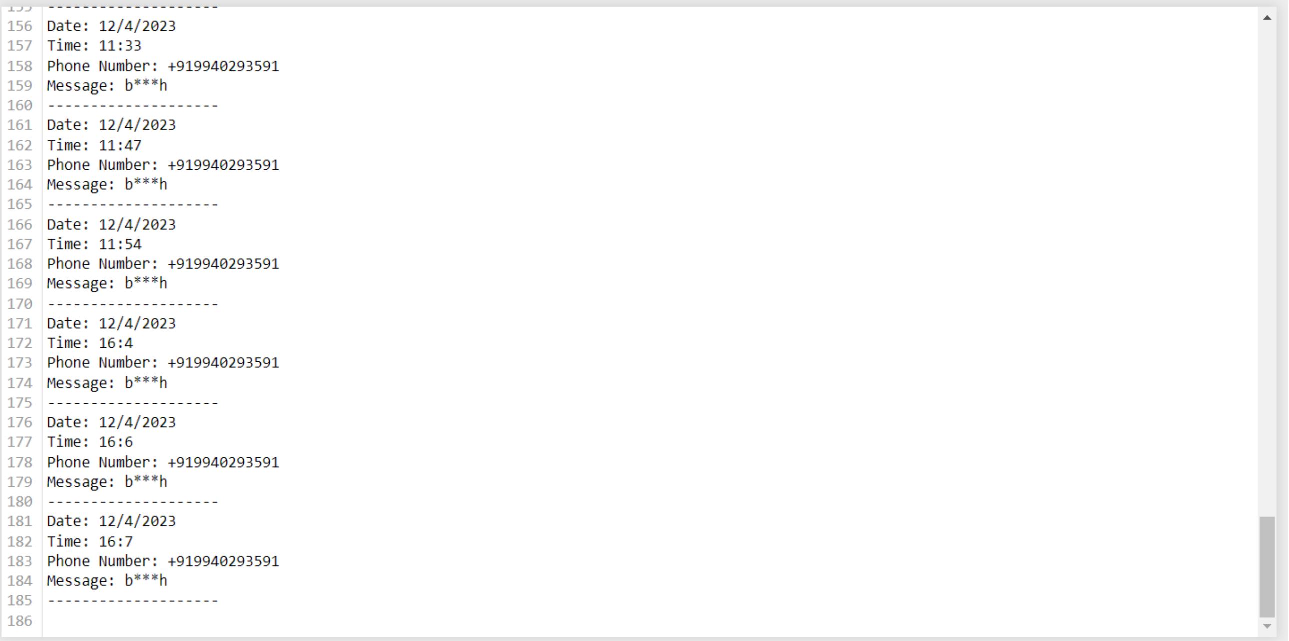
Sent from your Twilio trial account - {"Date": "25/03/2023", "Time": "<module 'time' (built-in)>", "Contact": "+91[redacted]91", "Message": "b[redacted]h", "Prediction": "abusive"}

Mar 25, 5:31 PM • Jio 4G

Text (Jio 4G)

Figure 7.3 Alert Notification Received by the User

DOI: 10.17148/IJARCCE.2023.124167



```
--------------------
Date: 12/4/2023
Time: 11:33
Phone Number: +919940293591
Message: b***h
--------------------
Date: 12/4/2023
Time: 11:47
Phone Number: +919940293591
Message: b***h
--------------------
Date: 12/4/2023
Time: 11:54
Phone Number: +919940293591
Message: b***h
--------------------
Date: 12/4/2023
Time: 16:4
Phone Number: +919940293591
Message: b***h
--------------------
Date: 12/4/2023
Time: 16:6
Phone Number: +919940293591
Message: b***h
--------------------
Date: 12/4/2023
Time: 16:7
Phone Number: +919940293591
Message: b***h
--------------------
```

Figure 7.4 All DBlog Messages Received by the User

VIII. CONCLUSION AND FUTURE WORKS:

In this project, we proposed a hybrid machine learning model to detect cyberbullying on the Twitter social media network using SVM and Logistic Regression. The model demonstrated the potential to identify and remove cyberbullying content on social media platforms. Our analysis showed that false positives were often caused by the model misinterpreting sarcastic or ironic tweets as cyberbullying, while false negatives were caused by the model's inability to recognize subtle forms of cyberbullying.

The proposed version works only on cyberbullying utilizing text of tweets. We couldn't perform the evaluation when it comes to the customers' behaviour. Besides, we purpose to classify and come across CB tweets in a real-time move. Also have proposed most effective for the Twitter dataset solely; different Social Media platforms (SMP) consisting of Instagram, Flickr, YouTube, Facebook, etc., need to be investigated so that you can come across the trend of cyberbullying.

REFERENCES

[1] A F. Mishna, M. Khoury-Kassabri, T. Gadalla, and J. Daciuk, ``Risk factors for involvement in cyber bullying: Victims, bullies and bully_victims," Children Youth Services Rev., vol. 34, no. 1, pp. 63_70, Jan. 2012, doi: 10.1016/j.childyouth.2011.08.032.
[2] K. Miller, ``Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limitedavailable redress," Southern California Interdiscipl. Law J., vol. 26, no. 2, p. 379, 2016.
[3] A. M. Vivolo-Kantor, B. N. Martell, K. M. Holland, and R. Westby, ``A systematic review and content analysis of bullying and cyber-bullying measurement strategies," Aggression Violent Behav., vol. 19, no. 4, pp. 423_434, Jul. 2014, doi: 10.1016/j.avb.2014.06.008.
[4] S. H. Sampasa-Kanyinga, P. Roumeliotis, and H. Xu, ``Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian school children," PLoS ONE, vol. 9, no. 7, Jul. 2014, Art. no. e102145, doi: 10.1371/journal.pone.0102145.
[5] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, ``Improving cyberbullying detection with user context," in Proc. Eur. Conf. Inf. Retr., in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Arti_cial Intelligence and Lecture Notes in Bioinformatics, vol. 7814, 2013, pp. 693_696.
[6] A. S. Srinath, H. Johnson, G. G. Dagher, and M. Long, ``BullyNet: Unmasking cyberbullies on social networks," IEEE Trans. Computat. Social Syst., vol. 8, no. 2, pp. 332_344, Apr. 2021, doi: 10.1109/TCSS.2021.3049232.

DOI: 10.17148/IJARCCE.2023.124167

[7] A. Agarwal, A. S. Chivukula, M. H. Bhuyan, T. Jan, B. Narayan, and M. Prasad, ``Identi_cation and classi_cation of cyberbullying posts: A recurrent neural network approach using under-sampling and class weighting,'' in Neural Information Processing (Communications in Computer and Information Science), vol. 1333, H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King, Eds. Cham, Switzerland: Springer, 2020, pp. 113_120.

[8] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, ``Machine learning and feature engineering-based study into sarcasmand irony classi_cation with application to cyberbullying detection,'' Inf. Process. Manage, vol. 58, no. 4, Jul. 2021, Art. no. 102600, doi: 10.1016/j.ipm.2021.102600.

[9] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A.Sharma, S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, and M. Masud, ``Nature-inspired-based approach for automated Cyberbullying classication on multimedia social networking,'' Math. Problems Eng., vol.2021, pp. 1_12, Feb. 2021, doi: 10.1155/2021/6644652.

[10] B. A. Talpur and D. O'Sullivan, ``Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter,'' Informatics, vol. 7, no. 4, p. 52, Nov. 2020, doi: 10.3390/informatics7040052.