International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01583

Cyberbullying Detection and Prevention On Social Media

Madhusudhanam M R

Department of Computer Science and Engineering
Malnad College of Engineering
Hassan, India
madhusudhanammr@gmail.com

Ajay Sarangam

Department of Computer Science and Engineering
Malnad College of Engineering
Hassan, India
ajjsarangam@gmail.com

Mayur A B

Department of Computer Science and Engineering
Malnad College of Engineering
Hassan, India
mayurab2206@gmail.com

Nayana R
Department of Computer Science and Engineering
Malnad College of Engineering
Assistant Professor
Hassan, India
nay@mcehassan.ac.in

Likith U S
Hassan, India
likithus2@gmail.com
Department of Computer Science and Engineering

International Scientific Journal of Engineering and Management ISSN: 2583-6129
Volume: 03 Issue: 04 | April – 2024 DOI: 10.55041/ISJEM01583
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

***Abstract*—Social media have become a very viable medium for communication, collaboration, exchange of information, knowl- edge, and ideas. However, due to anonymity preservation, the incidents of hate speech and cyberbullying have been diversified across the globe. Formulating solution strategies for automatic detection of cyber aggression and hate speech, varying from machine learning models with vast features to more complex deep neural network models and different social media platforms.**

To address this gap, we have performed extensive prepro- cessing for monitoring social media contents and in making cyberspace a secure and safer place for all segments of society.

***Index Terms*—cyberbullying,cyber aggression, Social media**

I. INTRODUCTION

Cyberbullying is aggressive, abusive behavior towards a person or a group of people, repeatedly hurting them by spreading offensive content or engaging in other forms of social violence through using digital media. The biggest chal- lenge is that cyberbullying can spread quickly. The depressivesymptoms and self-injuries are highest among the victims of Cyberbullying. It can take the form of sharing/ posting offensive video content or uploading violent images or sharing the pictures without permission of the owner etc. However, cyberbullying via textual content is far more common and found across most platforms.

The proliferation of online platforms has empowered indi- viduals to express their thoughts and opinions more openly
than ever before. Social media giants like Twitter and Face- book have seamlessly integrated into our daily routines, es- pecially resonating with teenagers. However, alongside the benefits, the surge in online activity has brought about neg- ative repercussions, particularly for adolescents who are in- creasingly exposed to various forms of cyberbullying. This pervasive issue manifests through influential social attacks, facilitated by the anonymity afforded by these platforms, al- lowing individuals to perpetrate unkind actions with impunity. Cyberbullying encompasses a spectrum of harmful behaviours, from aggressive content and harassment to hate speech and discrimination, inflicting severe psychological distress on its victims. Shockingly, studies indicate that a significant por- tion of the population, particularly adolescent girls, endure cyberbullying, with over 60 percentage of social media users in the US reporting exposure to such abuse. Consequently, cyberbullying has emerged as a global epidemic, necessitating concerted efforts to mitigate its impact and safeguard internet users' well-being. However, identifying offensive language poses a formidable challenge, given the nuances and variability inherent in human communication. Sarcasm, colloquialisms, and friendly banter often blur the lines between harmless interactions and malicious intent. Hence, automated detection systems have become imperative, spurring researchers world- wide to develop innovative models for distinguishing offensive from non-offensive content.

II. RELATED WORKS

Due to technology development, bullying is not only re- stricted to the physical and real life school context. Cyber- bullying refers to an intentional act of aggression, carried out to harm another individual using electronic forms of contacts or devices. Even though more and more research has linked cyberbullying to different aspects of psychological distress, there is still a lack of studies focused on how distinctive cyber- bullying plays. Cyberbullying detection approaches focusing on online social networks (OSNs) are reviewed. Divided cyberbullying occurrence into various themes, including race, sexuality, culture, and intelligence. Consequently, they utilizedsome controversial videos from YouTube as a use case to classify the comments posted on them using four different classifiers (Naive Bayes (NB), Rule-based Jrip, Tree-based J48, and SVM) [1].

Sanchez, H.; Kumar, S. Twitter bullying detection. Ser. NSDI 2011, 12, 15. Sanchez et al. Were one of the first to pro- pose a method to detect cyberbullying on the Twitter platform. The authors utilized the NB classifier to detect tweets that contained abusive behavior toward a specific gender. However, their method achieved only an accuracy of 70 percent and the size of the used dataset is relatively small [2].

Gamback, B.; Sikdar, U.K. Using convolutional neural networks to classify hatespeech. In Proceedings of the First Workshop on Abusive Language Online, August 2017; Associ- ation for Computational Linguistics: Vancouver, BC, Canada; presented a deep learning detection system to identify Twitter cyberbullying comments. This system classified the commentsinto one of four possible categories: sexism, racism, both (i.e., sexism and racism), and non-offensive comments [3].

Pradhan, A.; Yatam, V.M.; Bera, P. Self-attention for cyber- bullying detection. In Proceedings of the 2020 International Conference Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland, 15–19 June 2020; pp. 1–6. Examined the effectiveness of self-attention models (these models

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01583

achieved state-ofthe-art results in various ma- chine translation tasks) in cyberbullying detection. The authors

architecture using three data sources: Formspring, Wikipedia, and Twitter cyberbullying datasets [4].

Online social networks (OSNs) play an integral role in facilitating social interaction; however, these social networks increase antisocial behavior, such as cyberbul2 lying, hate speech, and trolling. Aggression or hate speech that takes placethrough short message service (SMS) or the Internet (e.g., in social media platforms) is known as cyberbullying. There- fore, automatic detection utilizing natural language processing (NLP) is a necessary first step that helps prevent cyberbullying[5].

The popularity of online social networks has created mas- sive social communication among their users and this leads to a huge amount of user-generated communication data. In recent years, Cyberbullying has grown into a major problem with the growth of online communication and social media. Cyberbullying has been recognized recently as a serious na- tional health issue among online social network users and developing an efficient detection model holds tremendous practical significance [6].

III. About Project

A. Problem Statement

Online threats and mean, aggressive, rude texts, tweets, posts, or messages are the root cause of cyberbullying. Despitehaving inbuilt systems to detect such issues, most social platforms lack accuracy in carrying out the task.

The escalated usage of social networking sites and freedom of speech has given optimal ground to individuals across all demographics for cyberbullying and cyber aggression. This leaves drastic and noticeable impacts on behavior of a victim, ranging from disturbance in emotional wellbeing and isolationfrom society to more severe and deadly consequences. Auto- matic Cyberbullying detection has remained a very challenging task since social media content is in natural language and is usually posted in unstructured free-text form leaving behind the language norms, rules, and standards.

Even though more and more research has linked cyberbully- ing to different aspects of psychological distress, there is still a lack of studies focused on how distinctive cyberbullying plays.

explored the usefulness of a self-attention model known as transformer

B. Objective

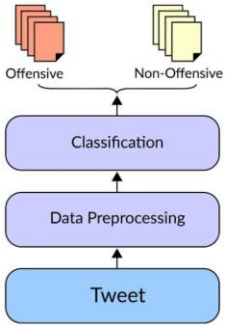
By building the scoring system for each user profile, the admin will be able to maintain a sentimental track record of each user allowing them to rank each user on the basis of theirscore.

The scoring is based on a periodic filtration process based on their profanity percentage. Each time the user is below the minimum requirement, their option to comment or post is disabled for hours.

Therefore, automatic detection utilizing natural language processing (NLP) is a necessary first step that helps prevent cy- berbullying. This research proposes an automatic cyberbully- ing method to detect aggressive behavior using a consolidated deep learning model. This technique utilizes multichannel deep learning based on three models, namely, the bidirectional gatedrecurrent unit (BiGRU), transformer block, and convolutional neural network (CNN), to classify Text based comments and chats into two categories: aggressive and not aggressive.

IV. PROJECT DESIGN

Convolution neural networks (aka CNN), originally incorpo-rated for image processing tasks, have become very efficaciousin different NLP and text classification applications. Thenetwork identifies correlations and patterns of data via theirfeature maps. Information about local structure of data is ex-tracted by applying multiple filters with different dimensions.Deep learning is a new discipline within the machine learn- ing field. Deep learning algorithms inspired by the structureand function of the brain are called artificial neural networks. Moreover, deep learning algorithms are trained to extract andunderstand the meaningful representations from the data itself



International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01583

Method	Accuracy
extra trees	65.62%
RF	69.29%
LR	54.27%
Bagging Classifier	68.30%
Linear SVC	50.33%

Fig. 1. The general framework of deep learning architecture

rather than performing straightforward classic programmed instructions. Meaningful representation is acquired by com- posing simple but non-linear modules that each transform the representation at one level (starting with the raw input data) into a representation at a higher level. Automatic detection of cyberbullying utilizing natural language processing (NLP) advanced machine learning methods on Online social networkssuch as Twitter platform is a necessary and an important. Moreover, the existing dataset that used to train the machine learning model are limited. Thus, it is very critical to have a large dataset that can cover many cyberbullying cases.

- Text cleaning : Text cleaning refers to the process of preparing textual data extracted from online platforms
- and contractions, cyberbullying detection systems can improve their ability to recognize and interpret potentially abusive language accurately.

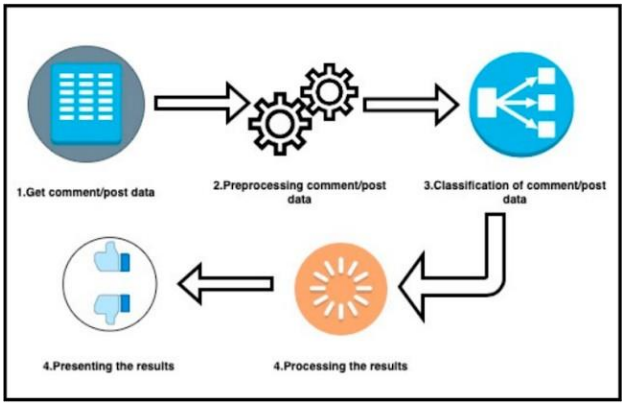


Fig. 2. The flow working of text based recognition of profanity

Three well-known hate speech datasets were combined to evaluate the performance of the proposed method. The

(such as social media posts, comments, messages, etc.) for analysis and detection of cyberbullying incidents.This process involves various techniques aimed at standardiz- ing, normalizing, and filtering the text to improve the accuracy and effectiveness of cyberbullying detection algorithms.

- Tokenization : Tokenization refers to the process of breaking down a piece of text, such as a social media post, comment, or message, into individual words, phrases, or tokens.This technique is commonly used as a prepro- cessing step in natural language processing (NLP) tasks, including cyberbullying detection, to facilitate further analysis and feature extraction.
- Filtering stop words : Filtering stop words in the con-text of cyberbullying involves removing common words from text data that do not carry significant meaning or contribute to the identification of cyberbullying behavior.
- Mapping slangs and contractions : Mapping slangs and contractions in the context of cyberbullying involves identifying and normalizing informal language and abbre- viated forms commonly used in online communication. This process is crucial for cyberbullying detection be- cause offenders often employ slang terms, abbreviations, and contractions to obscure their identity or convey harmful messages more discreetly. By mapping slangs

proposed method achieved good accuracy results.The multi-channel technique integrates the features of three deep learning models: transformer block, BiGRU and CNN. This integration helps to contribute to the final prediction.

```
1. IMPORT SL_RU
2. M_dic ← {}
3. For line in SL_RU Do:
4.    if not empty line:
5.       Append and expand items one by one
         Set key ← Slang, Value← Phrases/Terms
6.    End if
7. End for
8. PROCEDURE SlangMapping()
   For word in Tw Do:
   If word in M_dic:
         Retrieve value
   Else:
         append same word
   End For
   Return new Tw
9. END PROCEDURE
```

Fig. 3. Basic pseudo code for text approach

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01583

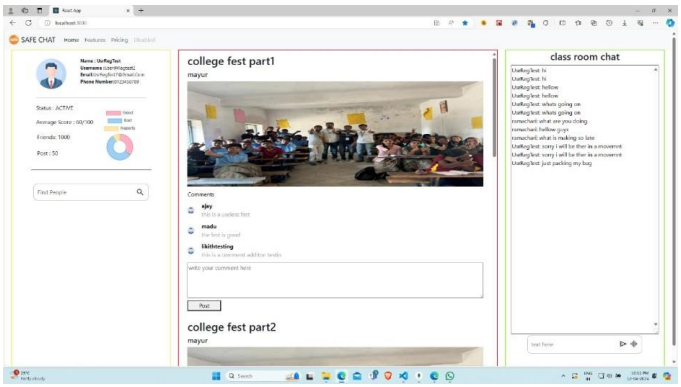


Fig. 4. Screen grab from the Main Webpage

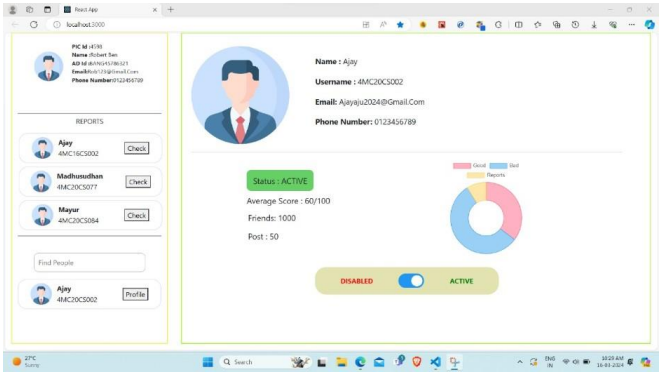


Fig. 5. Screen grab from the Student Analytics Webpage

V. IMPLEMENTATION

Social media have become a very viable medium for communication, collaboration, exchange of information, knowledge, and ideas. However, due to anonymity preservation, the incidents of hate speech and cyberbullying have been diversified across the globe. Formulating solution strategies for automatic detection of cyber aggression and hate speech, varying from machine learning models with vast features to more complex deep neural network models and different social media platforms.

To address this gap, we have performed extensive preprocessing for monitoring social media contents and in making cyberspace a secure and safer place for all segments of society.

Speech-to-Text has three main methods to perform speech recognition. These are listed below:

Synchronous Recognition (REST and gRPC) sends audio data to the Speech-to-Text API, performs recognition on that data, and returns results after all audio has been processed. Synchronous recognition requests are limited to audio data of 1 minute or less in duration.

Asynchronous Recognition (REST and gRPC) sends audio data to the Speech-to-Text API and initiates a Long

```
from better_profanity import profanity

# text to be censored
text = "Crap! What a stupid decision!"

# do censoring
censored = profanity.censor(text)

censored
```

```
'****! What a **** decision!'
```

Streaming Recognition (gRPC only) performs recognition on audio data provided within a gRPC bi-directional stream. Streaming requests are designed for real-time recognition purposes, such as capturing live audio from a microphone. Streaming recognition provides interim results while audio is being captured, allowing results to appear, for example, while a user is still speaking.

Requests contain configuration parameters as well as audio data. The following sections describe these types of recognition requests, the responses they generate, and how to handle those responses in more detail.

Speech-to-Text API recognition

A Speech-to-Text API synchronous recognition request is the simplest method for performing recognition on speech audio data. Speech-to-Text can process up to 1 minute of speech audio data sent in a synchronous request. After Speech-to-Text processes and recognizes all of the audio, it returns a response.

A synchronous request is blocking, meaning that Speech-to-Text must return a response before processing the next request. Speech-to-Text typically processes audio faster than realtime, processing 30 seconds of audio in 15 seconds on average. In cases of poor audio quality, your recognition request can take significantly longer.

Dealing with raw unstructured text data is a bit of a challenge. Conversational text as data presents itself with stray characters, punctuation, spellings, abbreviations, emojis etc to name a few. The presence of profanity in text data is the main focus of this article—ways to handle profanity in text and observing the impact that profanity/ censorship is likely to have on sentiment analysis.

Flag texts containing profanity, so that they can be filtered and passed through a data pipeline for a certain action to be taken such as warning the user or removing them from public view. As an alternative to completely removing swear words from text, those identified as inappropriate for plain view can

International Scientific Journal of Engineering and Management
Volume: 03 Issue: 04 | April – 2024
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM01583

be masked with the use of special characters. Many of us would have seen this widely used in subtitles on tv shows/ transcripts from virtual meetings.

The profanity.censor method is used to mask inappropriate words with special characters. The default setting uses "*".

In the context of cyberbullying detection, the profanity.censor method refers to a function or tool designed to detect and censor profane or offensive language in text data. This method is an essential component of cyberbullying detection systems as it helps identify and mitigate harmful content before it can cause harm to users. Here's how the profanity.censor method typically works:

- Profanity Detection: The method analyzes text data to identify words or phrases that are considered profane, offensive, or inappropriate. This can involve comparing the text against a predefined list of profanity words or using natural language processing techniques to detect offensive language based on linguistic patterns.
- Censorship: When profane or offensive language is detected, the method replaces or censors the offending words or phrases with non-offensive alternatives. This could involve replacing the profanity with asterisks (*) or other symbols, removing the offending words entirely, or replacing them with less offensive alternatives.
- Customization: The profanity.censor method may allow for customization to suit the specific needs and preferences of users or organizations. For example, users may have the option to add or remove words from the profanity list, adjust censorship settings, or define their own criteria for what constitutes offensive language.
- Integration: The method is typically integrated into cyberbullying detection systems or content moderation platforms to automatically scan and filter text data in real-time. This integration ensures that offensive content is promptly identified and addressed, helping to create safer online environments for users.

  Overall, the profanity.censor method plays a crucial role in cyberbullying detection by detecting and censoring profane or offensive language in text data, thereby helping to mitigate the impact of cyberbullying and create more respectful and inclusive online communities.

VI. RESULT

By integrating various filters through which profanity can be detected has allowed us to measure a score. This score allows us to rank each user and categorize him/her as a potential threat. With up to 80 percentage accuracy for Text recognition, and 75 percentage accuracy for voice recognition, we are able to accurately classify each user. This promotes the use of active safe chat features and blocks the user if found guilty; Forcing them to stop bullying, and engage is healthier conversations.

In order to instantly prevent the dissemination of vulgar texts on social media, we have implemented a system that integrates multiple filters to scrutinize the messages sent by users. This system identifies any words deemed vulgar or profane within the text and promptly masks them. By doing so, we aim to mitigate and prevent instances of cyberbullying.

```
from profanity_check import predict, predict_prob

predict_prob(['Crap!What a stupid decision!'])

array([0.8683693])
```

86% — probability of containing profanity

```
from profanity_check import predict, predict_prob

predict_prob(['Lovely spring weather!'])

array([0.06606084])
```

6.6% probability of containing profanity

Fig. 7. Profanity from Text-Analysis

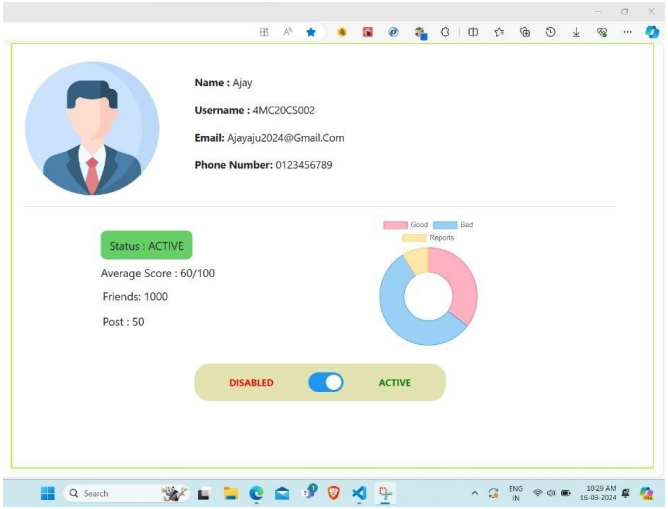


Fig. 8. Score Allocation

This proactive approach not only reduces the spread of harmful language but also fosters a safer and more respectful online environment.

VII. CONCLUSION

International Scientific Journal of Engineering and Management ISSN: 2583-6129
Volume: 03 Issue: 04 | April – 2024 DOI: 10.55041/ISJEM01583
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Due to technology development, bullying is not only restricted to the physical and real-life school context. Cyberbul- lying refers to an intentional act of aggression, carried out to harm another individual using electronic forms of contacts or devices.

By building the scoring system for each user profile, the admin will be able to maintain a sentimental track record of each user allowing them to rank each user on the basis of their score.

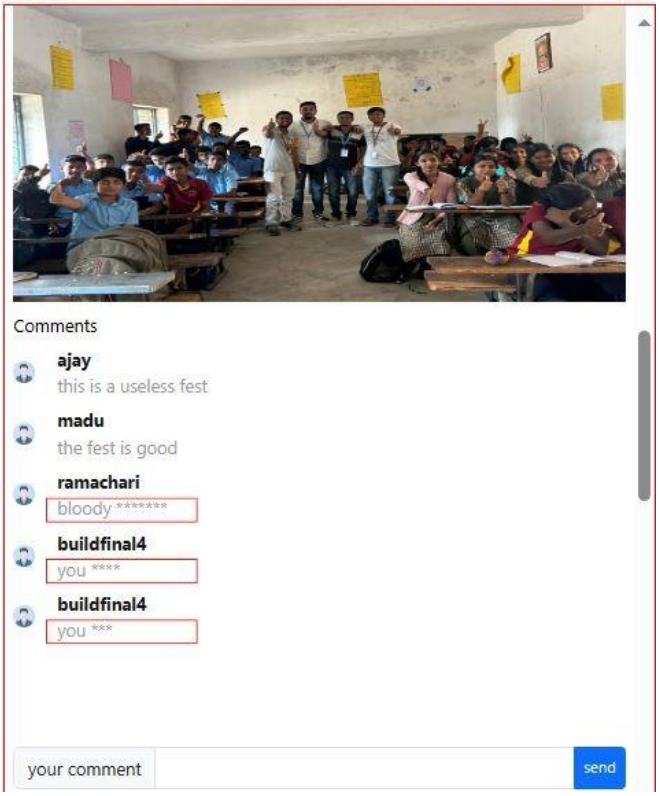


Fig. 9. Masked texts

The scoring is based on a periodic filtration process based on their profanity percentage. Each time the user is below the minimum requirement, their option to comment or post is disabled for hours.

Cyberbullying has become an alarming social threat for today's youth and has recently gained huge attention from the research community. This research has addressed the problem of cyberbullying detection in cyberbullying has become an alarming social threat for today's youth and has recently gained huge attention from the research community. This research has addressed the problem of cyberbullying detection and prevention by providing a ranking mechanism.

The outcomes of this study, if implemented, will assist cybercrime centers and investigation agencies for monitoring social media contents and in making cyberspace a secure and safer place for all segments of society.

ACKNOWLEDGMENT

We present with immense pleasure this work titled "Cyberbullying Detection and prevention on Social media ". An endeavour over a long period can be successful with the advice and support of many well wishers. We take this opportunity to express our gratitude and appreciation toall of them. The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without mentioning the people who made it possible. So, with gratitude we acknowledge all those whose guidance and encouragement made to suc- cessfully complete this project. We would like to express sincere thanks to our Principals Dr.A.J. Krishnaiah, Malnad College of Engineering for his encouragement made to successfully completion of the project work. We wish to express our gratitude to Dr.Geetha Kiran A, Professor and Head, Department of Computer Science Engineering for providing a good working environment and for her constant support and encouragement. It gives great pleasure in placing on record a deep sense of gratitude toour guide Mrs.Nayana R,Assistent Professor , Department of Computer Science Engineer- ing for his daily evaluation of the work and for providing us constant encouragement withhis unflinching support and valuable guidance throughout this project. We would also like to thank all the staff of Computer Science and Engineering department who have directly orindirectly helped us in the completion of the project work. At last we would hereby acknowledge and thank our parents who have been a source of inspiration and also instrumental in the successful completion of the project work

REFERENCES

[1] Hellfeldt K, López-Romero L, Andershed H. Cyberbullying and psychological well-being in young adolescence: the potential protective mediation effects of social support from family, friends, and teachers. Int J Environ Res Public Health. 2020;17(1):45.

[2] Alotaibi M, Alotaibi B, Razaque A. A multichannel deep learning framework for cyberbullying detection on social media. Electronics. 2021;10(21):2664.

[3] Gamback, B.; Sikdar, U.K. Using convolutional neural networks to classify hate-speech. In Proceedings of the First Workshop on Abusive Language Online, August 2017; Association for Computational Linguistics: Vancouver, BC, Canada;

[4] Pradhan, A.; Yatam, V.M.; Bera, P. Self-attention for cyberbullying detection. In Proceedings of the 2020 International Conference Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland, 15–19 June 2020; pp. 1–6.

[5] A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media by Munif Alotaibi 1,*ORCID,Bandar Alotaibi 2,*ORCID andAbdul Razaque, Abdul Razaque. Electronics 2021, 10(21), 2664;

[6] Al-garadi, M.A.; Varathan, K.D.; Ravana, S.D. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Comput. Hum. Behav. 2016, 63, 433–443.