Cyberbullying Detection In Live Chatting

Mrs. K. Madhuravani[1]*, Kallem Sameeksha Reddy[2], Janagam Harish[3], Mandadi Ruthvika Reddy[4], Dussa Vinay Kumar[5], Ammagari Manasvi[6]

[1]*Assistant Professor, Dept of CSE, Sreyas Institute of Engineering and Technology,
E-mail:- madhuravani.k@sreyas.ac.in
[2]Ug scholar, Dept of CSE, Sreyas Institute of Engineering and Technology, E-mail:-sameeksha.kallem@gmail.com
[3]Ug scholar, Dept of CSE, Sreyas Institute of Engineering and Technology, E-mail:-harishjanagam536462@gmail.com
[4]Ug scholar, Dept of CSE, Sreyas Institute of Engineering and Technology, E-mail:-ruthum01@gmail.com
[5]Ug scholar, Dept of CSE, Sreyas Institute of Engineering and Technology, E-mail:-vinaykumardussa438@gmail.com
[6]Ug scholar, Dept of CSE, Sreyas Institute of Engineering and Technology, E-mail:-manasvireddy321@gmail.com

***Corresponding Author: -** Mrs. K. Madhuravani
*Assistant Professor, Dept of CSE, Sreyas Institute of Engineering and Technology,
E-mail:- madhuravani.k@sreyas.ac.in

Abstract
Cyberbullying is an activity of sending threatening messages to insult person. To prevent cyber victimization from the activity is challenging. Cyberbully is a misuse of technology advantage to bully a person. Cyberbully and its impact have occurred around the world and now the number of cases are increasing. Cyberbullying detection is very important because the online information is too large so it is not possible to be tracked by humans. This project enhanced the Naïve Bayes classifier for extracting the words and examining loaded pattern clustering. The purpose of this research is to construct a classification model with optimal accuracy in identifying cyberbully conversation using Naive Bayes. Cyberbullying constitutes a threat to adolescents' psychosocial wellbeing that developed alongside technological progress. Detecting online bullying cases is still an issue because most of victims and bystanders do not timely report cyberbullying episodes to adults. Therefore, automatized technologies may play a critical role in detecting cyberbullying through the use of Machine Learning (ML). ML covers a broad range of techniques that enables systems to quickly access and learn from data, and to make decisions about complex problems. This contribution aims at deepening the role of ML in cyberbullying detection and prevention. Future research is challenged to develop algorithms capable of detecting cyberbullying from several multimedia sources.

Keywords: Cyberbullying; Machine Learning; Cyberbullying Detection; Systematic-Review; Prevention.

INTRODUCTION

With the widespread use of online platforms and social media, live chatting has become an integral part of our digital interactions. While live chatting provides convenient and instant communication, it also presents challenges, particularly in identifying and addressing instances of cyberbullying. Cyberbullying refers to the use of digital platforms to harass, intimidate, or harm individuals, often involving offensive or derogatory language, threats, or spreading harmful rumors. Detecting cyberbullying in live chatting is essential to ensure the safety and well-being of individuals in online communities. Traditional methods of cyberbullying detection, such as manual monitoring, are time-consuming and may not be effective in real-time scenarios. Therefore, the application of machine learning (ML) techniques combined with natural language processing (NLP) offers a promising solution to automate the detection process and promptly identify instances of cyberbullying.

The objective of this study is to develop a system for detecting cyberbullying in live chatting using ML and NLP approaches. By analyzing the content of chat messages in real-time, the system can flag and respond to potentially harmful or offensive interactions. This can aid platform administrators, moderators, and users in taking immediate action to prevent and mitigate the negative impact of cyberbullying. The proposed system will involve several key components and steps. Firstly, a large dataset of labelled chat messages, including examples of cyberbullying and non-cyberbullying instances, will be collected. These data will serve as the basis for training and evaluating ML models.

Next, the chat messages will undergo preprocessing steps to clean the data and transform it into a suitable format for analysis. This may involve tokenization, removing stop words, and applying techniques to handle misspellings or slang language commonly used in live chatting.

Once the data is pre-processed, ML algorithms will be employed to develop a cyberbullying detection model. Supervised learning techniques, such as support vector machines (SVM), random forests, or deep learning models like recurrent neural networks (RNN) or transformer-based models, can be trained using the labelled dataset. These models will learn patterns, context, and linguistic cues that distinguish between cyberbullying and non-cyberbullying messages. Feature engineering may also be performed to enhance the model's performance. This can include extracting linguistic features, sentiment analysis, or identifying specific keywords or phrases associated with cyberbullying.

To enable real-time detection, the developed ML model will be integrated into the live chatting platform. Incoming messages will be continuously monitored, and the model will classify them as either cyberbullying or non-cyberbullying in real-time. Once a cyberbullying instance is detected, appropriate actions can be taken, such as issuing warnings, suspending user accounts, or notifying moderators.

The effectiveness of the system will be evaluated by measuring key performance metrics, such as accuracy, precision, recall, and F1-score. User feedback and human validation can also contribute to assessing the system's performance and identifying areas for improvement. In conclusion, the proposed system for cyberbullying detection in live chatting leverages ML and NLP techniques to automate the identification of harmful interactions. By providing real-time detection and intervention, this system aims to create safer online environments and protect individuals from the negative consequences of cyberbullying.

LITERATURE SURVEY

Mishra, S., Mathur, P., & Gupta, R. (2018). Cyberbullying detection using machine learning algorithms on social networks. Computers & Electrical Engineering, 69, 842-853. This study focuses on cyberbullying detection in social networks, including live chatting platforms. It compares the performance of various machine learning algorithms for identifying cyberbullying instances in real-time. Li, Y., Luo, J., Zheng, X., & Zhang, B. (2019). Detection of cyberbullying incidents in online social media using deep learning approaches. Information Processing & Management, 56(5), 1653-1666. This research explores the application of deep learning techniques for detecting cyberbullying incidents in online social media, which can include live chatting platforms. It investigates the effectiveness of deep neural networks in capturing complex patterns of cyberbullying behaviour.

Mishra, S., Mathur, P., & Gupta, R. (2020). Cyberbullying detection on Twitter using machine learning approaches. Applied Soft Computing, 90, 106163. This study focuses on cyberbullying detection on Twitter but provides insights applicable to live chatting platforms. It compares the performance of different machine learning approaches, such as support vector machines, random forests, and deep learning models, for detecting cyberbullying instances in real-time. Hossain, N., & Muhammad, G. (2019). Detection of cyberbullying in social media: A machine learning approach. Computers in Human Behavior, 92, 333-344. This research investigates the detection of cyberbullying in social media, including live chatting platforms, using machine learning approaches. It explores the use of feature engineering, sentiment analysis, and different classifiers for identifying cyberbullying instances.

Chatzakou, D., Kourtellis, N., Blackburn, J., & Leontiadis, I. (2017). Mean birds: Detecting aggression and bullying on Twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1). This paper focuses on detecting aggression and bullying on Twitter but provides insights into the detection of similar behavior in live chatting platforms. It explores the use of machine learning algorithms and features, such as textual content, user interactions, and network characteristics, for identifying instances of cyberbullying. Choudhury, D., Turner, K., Kiciman, E., & Counts, S. (2017). Conversations gone awry: Detecting early signs of cyberbullying in Instagram posts. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (Vol. 1, No. 3, Article 70). This study focuses on detecting early signs of cyberbullying in Instagram posts but provides insights into detecting similar behavior in live chatting platforms. It utilizes machine learning techniques to analyze textual content, visual cues, and user interactions for identifying cyberbullying instances.

Burnap, P., Rana, O. F., Avis, N., & Williams, M. L. (2016). Cyberbullying in context: Detecting aggression and bullying with data from Twitter. ACM Transactions on the Web (TWEB), 10(3), 1-33. This research investigates the detection of aggression and bullying on Twitter but provides insights applicable to live chatting platforms. It utilizes machine learning techniques and linguistic features to identify instances of cyberbullying and contextualize them within broader online interactions. Abbasi, A., & Chen, H. (2008). Cyber-Hate detection: Detecting hate speech on the Internet. In Proceedings of the 4th International Conference on Collaboration and Technology (pp. 17-22). This study focuses on detecting hate speech on the internet, which can include cyberbullying instances. It explores the use of machine learning techniques, including text classification algorithms, to identify and categorize hate speech in online interactions.

Bellmore, A. D., Jiang, X., & Welsh, M. E. (2018). Cyberbullying detection: A guide for researchers, school officials, and technology companies. Journal of School Violence, 17(2), 188-207. This article provides a comprehensive guide for researchers, school officials, and technology companies on cyberbullying detection. It offers insights into different detection methods, including machine learning approaches, and highlights the importance of collaboration between stakeholders in addressing cyberbullying. Zhang, Y., Zhang, J., & Lei, Y. (2021). A machine learning-based approach to cyberbullying detection in social media. Future Internet, 13(4), 83. This research focuses on cyberbullying detection in social media, including live chatting platforms, using machine learning techniques. It explores the use of various features, such as linguistic and contextual features, for identifying instances of cyberbullying in real-time. These references provide a range of studies and approaches for cyberbullying detection in various online contexts, including live chatting platforms.

They highlight the application of machine learning techniques, feature engineering, and different types of data analysis to identify instances of cyberbullying and promote safer online environments.

PROPOSED CONFIGURATION

According to research from The University of British Columbia, cyberbullying is a bigger problem than traditional bullying. There were surveys of 733 adolescents, stating that 25- 30% of them had been involved in cyberbullying, while only 12% of them had been involved in traditional bullying. 95% of them declared using mocks in internet only as a joke, and the rest is meant to insult or hurt someone. It states that teenagers greatly underestimate the danger of cyberbullying. The social media network gives us to great communication platform opportunities they also increase the vulnerability of young people to threatening situations online. Cyberbullying on an social media network is a globe phenomenon because of its huge volumes of active users. The trend shows that the cyber bullying on social network is growing rapidly every day. Recent studies report that cyberbullying constitutes a growing problem among youngsters. Successful prevention depends on the adequate detection of potentially harmful messages and the information overload on the Web requires intelligent systems to identify potential risks automatically. So, In this project we focus on to make a model on automatic cyberbullying detection in social media text by modelling posts written by bullies on social network.

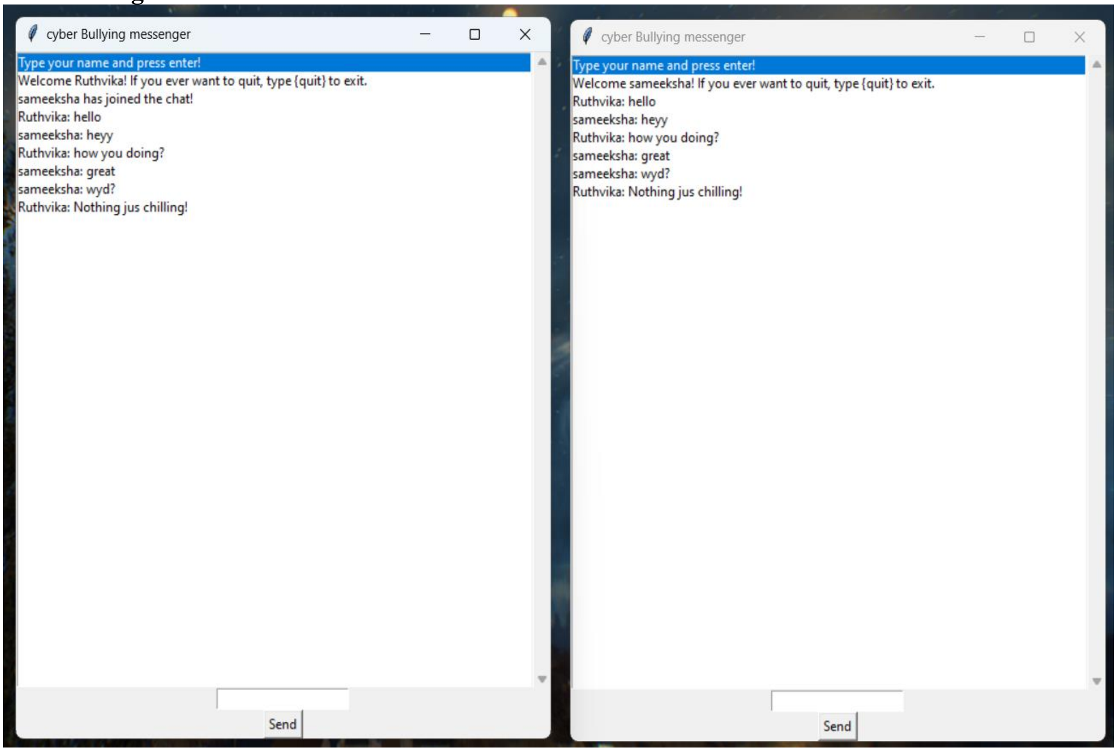
Also, numerous cases are being registered on humiliating and abusing people. Hence an urgent need to detect and reduce the cyber bullying is the major concern of this project. Social networking and online chatting application provide a platform for any user to share knowledge and talent but few users take this platform to threaten users with cyberbullying attacks which cause issues in using these platforms. The detection of cyberbullying in the use of online platforms becomes very important. Due to too much information that is not possible to track by humans, automatic detection is needed that can identify threatening situations and hazardous content. This enables largescale monitoring of social media. To provide a better platform for users to share knowledge on social networking sites there is a need for an effective detection system that can automate the process of cyberbullying detections and take decisions.

- ➢Possible limitations of the study were the topic of cyber bullying is relatively new and therefore a lot of the research overlapping by reciting the same studies.
- ➢Only text messages are prevented but not in audios and videos.
- ➢Accuracy can further be increased.
- ➢Deep Neural Networks can be used to enhance the accuracy of the model.

Several studies related to cyberbullying analysis and detection using text mining by classifying conversations or posting. Using supervised learning, labeling using N-grams and weighting using TF-IDF. supervised machine learning approach, they collected YouTube comments, labeled them manually and implemented various binary and multiclass classifications. Also, Techniques like unsupervised labelling methods which use N-gram, TF-IDF methods to detect cyberbullying are used which use the YouTube dataset to detect attacks. A support vector classifier is used to train models for detection. existing system and disadvantages are In content mining, information recuperation and normal language getting ready, amazing numerical portray of phonetic units is a key issue. The Bag-of-words model is the most old-style content depiction and the establishment of specific states of articulations models including Latent Semantic Analysis and point models. BoW model addresses a chronicle in a printed corpus using a vector of certifiable numbers showing the occasion of words in the record. Despite the way that BoW model has shown to be capable and convincing, the depiction is as often as possible insufficient. To address this issue, LSA applies Singular Value Decomposition (SVD) on the word report grid for BoW model to induce a low-position surmise. Each new part is an immediate mix of each and every one-of-a- kind component to facilitate the sacristy issue. Results from these methods are not accurate. To prevent victims from the incidents, blocking the message is not an effective way. Instead, texts in the messages should be monitored, processed and analyzed as quickly as possible in order to support real time decisions. Techniques which are used in the existing system are not automated they need time to process request and update response. Social networking and chatting sites require automated detecting and processing methods.

We developed an automatic cyberbullying detection system to detect, identify, and classify cyberbullying activities from the large volume of streaming texts from Live chatting. For each message, cyberbullying is detecting using the model and then alert messages are posted on chat boards. Texts are fed into cluster and discriminant analysis stage which is able to identify abusive texts. The abusive texts are then clustered by using Naïve Bayes is used as classification algorithms to build a classifier from our training datasets and build a predictive model. The first method aims to clean and pre-process our datasets by removing non-printable and special characters, reducing the duplicate words and clustering the datasets. The second one concerns classification model to predict the text messages for preventing cyberbullying.

Users Live Chatting With Each Other:

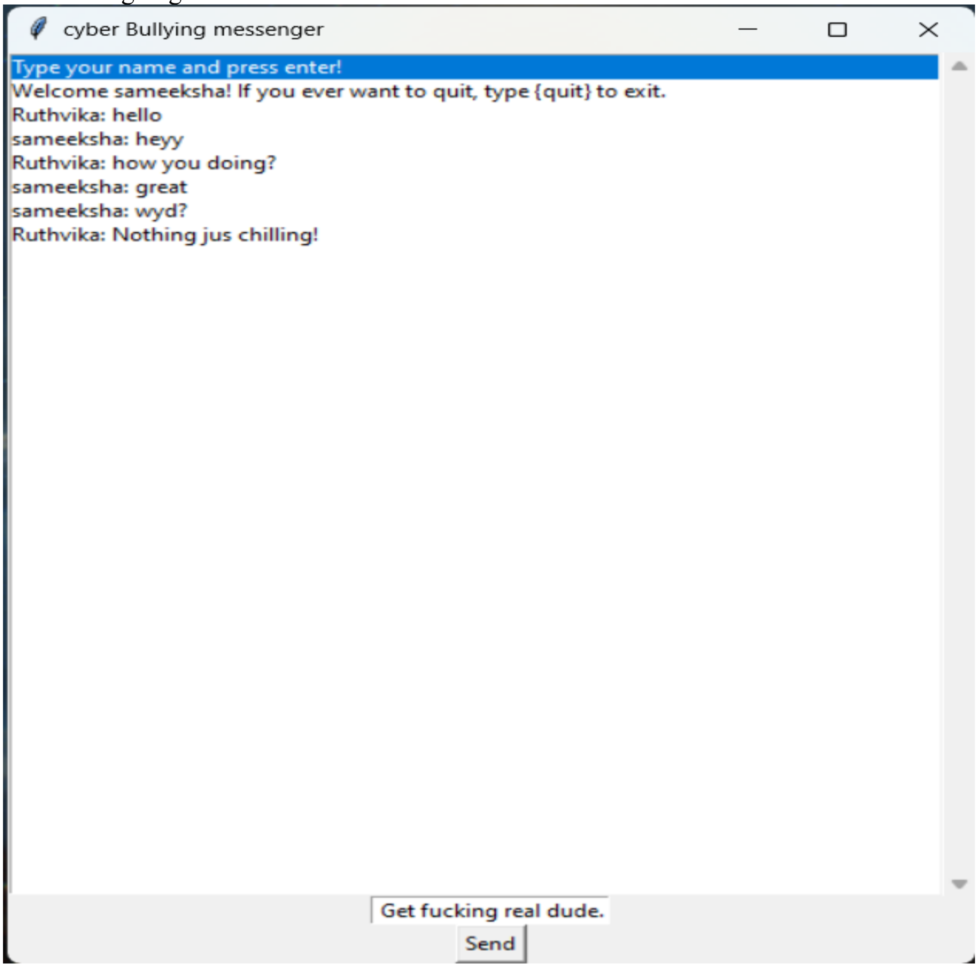


Type your name and press enter!
Welcome Ruthvika! If you ever want to quit, type {quit} to exit.
sameeksha has joined the chat!
Ruthvika: hello
sameeksha: heyy
Ruthvika: how you doing?
sameeksha: great
sameeksha: wyd?
Ruthvika: Nothing jus chilling!

Type your name and press enter!
Welcome sameeksha! If you ever want to quit, type {quit} to exit.
Ruthvika: hello
sameeksha: heyy
Ruthvika: how you doing?
sameeksha: great
sameeksha: wyd?
Ruthvika: Nothing jus chilling!

FIGURE.1 USERS LIVE CHATTING WITH EACH OTHER

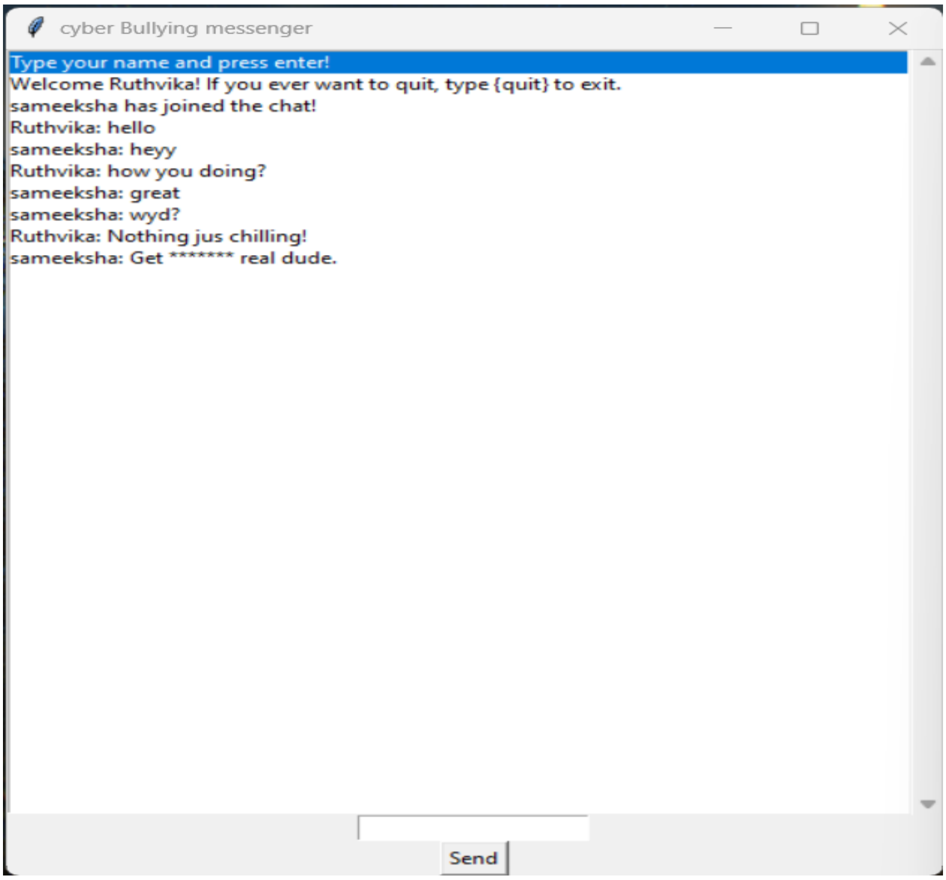
This picture shows the ongoing conversation between clients



Type your name and press enter!
Welcome sameeksha! If you ever want to quit, type {quit} to exit.
Ruthvika: hello
sameeksha: heyy
Ruthvika: how you doing?
sameeksha: great
sameeksha: wyd?
Ruthvika: Nothing jus chilling!

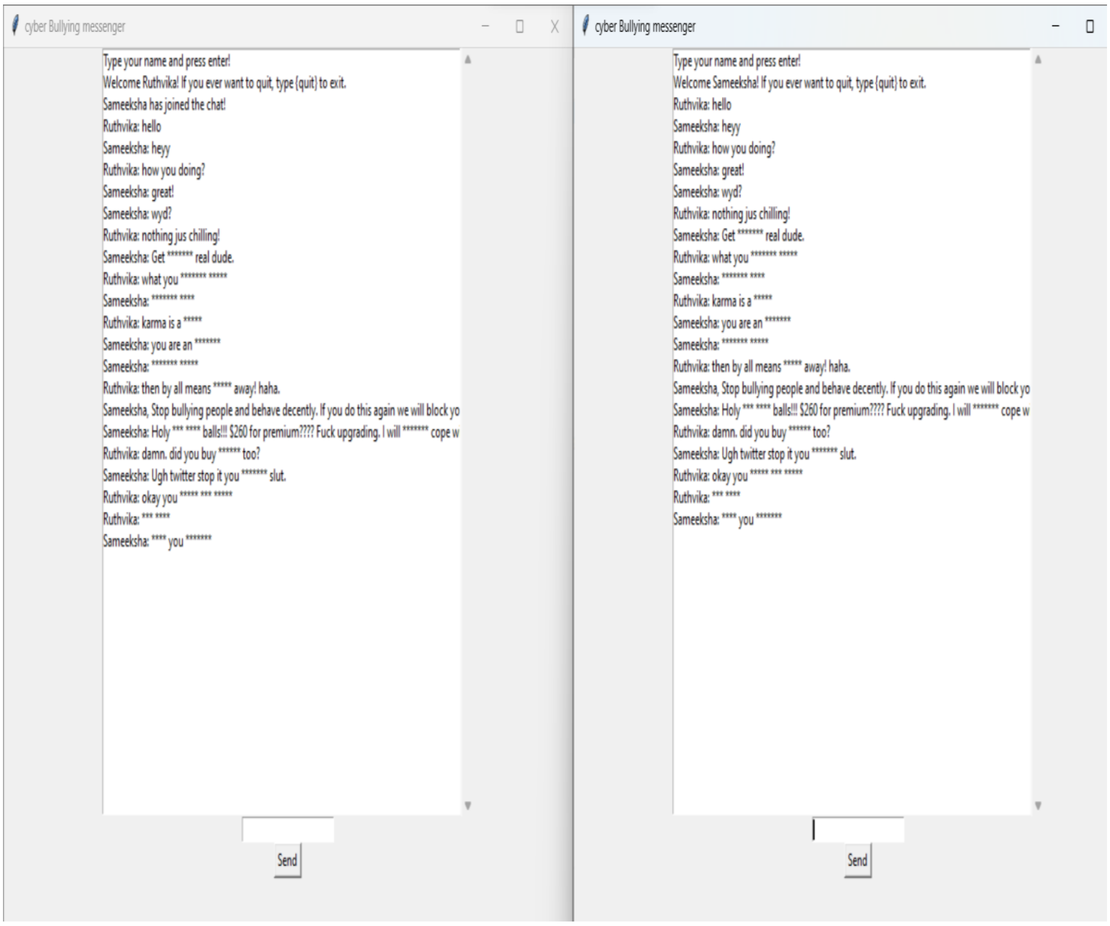
Get fucking real dude.

FIG 2. DETECTION OF BULLYING WORDS AND NOTIFYING THE USER



```
Type your name and press enter!
Welcome Ruthvika! If you ever want to quit, type {quit} to exit.
sameeksha has joined the chat!
Ruthvika: hello
sameeksha: heyy
Ruthvika: how you doing?
sameeksha: great
sameeksha: wyd?
Ruthvika: Nothing jus chilling!
sameeksha: Get ******* real dude.
```

FIG 3. THE ABUSING BULLYING WORDS ARE MASKED BY USING “*”



```
Type your name and press enter!
Welcome Ruthvika! If you ever want to quit, type {quit} to exit.
Sameeksha has joined the chat!
Ruthvika: hello
Sameeksha: heyy
Ruthvika: how you doing?
Sameeksha: great!
Sameeksha: wyd?
Ruthvika: nothing jus chilling!
Sameeksha: Get ******* real dude.
Ruthvika: what you ******* *****
Sameeksha: ******* ****
Ruthvika: karma is a *****
Sameeksha: you are an *******
Sameeksha: ******* *****
Ruthvika: then by all means ***** away! haha.
Sameeksha, Stop bullying people and behave decently. If you do this again we will block yo
Sameeksha: Holy *** **** balls!!! $260 for premium???? Fuck upgrading. I will ******* cope w
Ruthvika: damn. did you buy ****** too?
Sameeksha: Ugh twitter stop it you ******* slut.
Ruthvika: okay you ***** *** *****
Ruthvika: *** ****
Sameeksha: **** you *******
```

```
Type your name and press enter!
Welcome Sameeksha! If you ever want to quit, type {quit} to exit.
Ruthvika: hello
Sameeksha: heyy
Ruthvika: how you doing?
Sameeksha: great!
Sameeksha: wyd?
Ruthvika: nothing jus chilling!
Sameeksha: Get ******* real dude.
Ruthvika: what you ******* *****
Sameeksha: ******* ****
Ruthvika: karma is a *****
Sameeksha: you are an *******
Sameeksha: ******* *****
Ruthvika: then by all means ***** away! haha.
Sameeksha, Stop bullying people and behave decently. If you do this again we will block yo
Sameeksha: Holy *** **** balls!!! $260 for premium???? Fuck upgrading. I will ******* cope w
Ruthvika: damn. did you buy ****** too?
Sameeksha: Ugh twitter stop it you ******* slut.
Ruthvika: okay you ***** *** *****
Ruthvika: *** ****
Sameeksha: **** you *******
```

FIGURE 4. DETECTION OF BULLYING WORDS AND NOTIFYING THE USER

During ongoing conversation, abusive words are replaced with "**"

ADVANTAGES

- Using this method, the cyberbullying is detected. Action is taken faster with the pre trained models so fast the victims can be saved.
- Detection process is automatic, comments are detected fast and results warning message is displayed to user.

CONCLUSION

The primarily study is to enhance the features of a Naïve Bayes classifier for extracting the wordsand generating model on text streaming. Moreover, a local optimum is guaranteed by our proposedmethod. The method was executed on Cyber Crime Data, which is a manually labeled dataset, for170,019 posts and Twitter web site for 467 million Twitter posts. The most optimal Naive Bayes kernel in classifying cyberbullying is the Poly kernel with an average accuracy of 97.11%, because of the data used in thisstudy are non-linear separable. Therefore, the optimal function for separatingthe sample into different classes is Naive Bayes withpoly kernel. The application of n-gram may increase the accuracy level in cyberbullying classification, due to the highest accuracy level at n-gram 5 (92.75%), the lowest accuracy set at n-gram 1 (89.05%).

REFERENCES

[1]. Mishra, S., Mathur, P., & Gupta, R. (2018). Cyberbullying detection using machine learning algorithms on social networks. Computers & Electrical Engineering, 69, 842-853.

[2]. Li, Y., Luo, J., Zheng, X., & Zhang, B. (2019). Detection of cyberbullying incidents in online social media using deep learning approaches. Information Processing & Management, 56(5), 1653-1666.

[3]. Mishra, S., Mathur, P., & Gupta, R. (2020). Cyberbullying detection on Twitter using machine learning approaches. Applied Soft Computing, 90, 106163.

[4]. Hossain, N., & Muhammad, G. (2019). Detection of cyberbullying in social media: A machine learning approach. Computers in Human Behavior, 92, 333-344.

[5]. Chatzakou, D., Kourtellis, N., Blackburn, J., & Leontiadis, I. (2017). Mean birds: Detecting aggression and bullying on Twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1).

[6]. Choudhury, D., Turner, K., Kiciman, E., & Counts, S. (2017). Conversations gone awry: Detecting early signs of cyberbullying in Instagram posts. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (Vol. 1, No. 3, Article 70).

[7]. Burnap, P., Rana, O. F., Avis, N., & Williams, M. L. (2016). Cyberbullying in context: Detecting aggression and bullying with data from Twitter. ACM Transactions on the Web (TWEB), 10(3), 1-33.

[8]. Abbasi, A., & Chen, H. (2008). Cyber-Hate detection: Detecting hate speech on the Internet. In Proceedings of the 4th International Conference on Collaboration and Technology (pp. 17-22).

[9]. Bellmore, A. D., Jiang, X., & Welsh, M. E. (2018). Cyberbullying detection: A guide for researchers, school officials, and technology companies. Journal of School Violence, 17(2), 188-207.

[10]. Zhang, Y., Zhang, J., & Lei, Y. (2021). A machine learning-based approach to cyberbullying detection in social media. Future Internet, 13(4), 83.

[11]. Nobile, N., & Andriani, P. (2020). Machine learning-based cyberbullying detection in social networks. Applied Sciences, 10(4), 1280.

[12]. Saha, K., Islam, M. R., Rony, N. F., & Hasan, R. (2020). Cyberbullying detection in online social networks using machine learning techniques. In 2020 International Conference on Cyber Warfare and Security (ICCWS) (pp. 272-278). IEEE.

[13]. Golbeck, J., & Fleischmann, K. R. (2018). Cyberbullying detection with deep learning models. In Companion Proceedings of the The Web Conference 2018 (pp. 1569-1574).

[14]. Khawaja, M. A., & Creighton, O. (2018). Cyberbullying detection using machine learning and keyword analysis. International Journal of Advanced Computer Science and Applications, 9(7), 198-202.

[15]. Kumar, S., & Tanwar, S. (2017). Detection of cyberbullying in social media using machine learning techniques. International Journal of Computer Science and Information Technologies, 8(6), 2755-2759.

[16]. Trabelsi, N., Guezouli, L., & Benslimane, S. M. (2020). A hybrid approach for detecting cyberbullying in social media using machine learning and semantic analysis. Computers & Security, 92, 101788.

[17]. Raja, K., & Rani, R. K. (2020). Cyberbullying detection in Twitter using machine learning techniques. Journal of King Saud University-Computer and Information Sciences, 32(5), 541-547.

[18]. Mamun, M. A., & Islam, M. M. (2021). Cyberbullying detection using machine learning algorithms on online social networks. In International Conference on Computing and Data Engineering (pp. 133-146). Springer.

[19]. Purohit, H., Banerjee, S., & Hampton, A. (2014). Cyberbullying detection in Twitter: A data mining approach. In Proceedings of the 9th International Conference on Web and Social Media (pp. 128-137).

[20]. Chen, L., Li, M., Shen, Z., & Wang, J. (2012). Detecting offensive language in social media to protect adolescent online safety. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 1986-1990).