Cyberbullying Detection Model Using AI&ML

Omkar K[1], Sakshi D[2], Priti D[3], Siddhi T[4] and Tushar S[5]
[1]*BE Student, CSE, Keystone School of Engineering*
[2]*Assistant Professor, CSE, Keystone School of Engineering*

***Abstract*—The popularity of social media, and other forms of online interaction, cyberbullying has become a huge problem, and the negative effects on the mental and emotional state of the user can be severe. As user generated content grows, it is hard to monitor every action for negative behavior. This research develops a cyberbullying detection system through machine learning. The proposed model can quickly and automatically identify instances of bullying and similar behavior in online text. This is achieved by applying natural language processing techniques to identify linguistic patterns and contexts associated with cyberbullying behavior. A labeled dataset containing text samples of bullying and non-bullying text was used to create the training and evaluation dataset. The text was preprocessed through tokenization, removal of stop words, and normalization. The model was developed in Python. The TF-IDF method was used to vectorize the text, and supervised learning approaches for text classification were the Naïve Bayes classifier and Support Vector Machines (SVM). The experimental results show that the proposed approach is effective and yields an accurate cyberbullying classification. The model can be further used to support social media platforms in the automatic content moderation to reduce the harmful interactions and create a safer user experience.**

***Index Terms*—Automated cyberbullying analysis, linguistic pattern analysis, Natural language processing, Social media.**

I. INTRODUCTION

The quick growth of the internet and social media has changed how people talk to and interact with each other. Sites like social networks, discussion forums, and messaging apps let users share opinions, ideas, and experiences right away. However, this ease of communication has also led to negative behaviors like cyberbullying. Cyberbullying is when people use digital platforms to harass, threaten, or insult others. It can have serious psychological, emotional, and social effects on victims, especially teenagers and young adults. With the growing amount of user-generated content on social media, it has become very difficult for platform administrators to manually monitor and control harmful content. Millions of posts, comments, and messages appear every day. This makes manual moderation inefficient and time-consuming. As a result, there is a strong need for automated systems that can identify and filter cyberbullying content effectively and efficiently. Machine learning techniques offer a promising way to automatically detect cyberbullying in online text. By examining patterns in language and user behavior, machine learning models can learn to tell the difference between normal communication and abusive or bullying content. Natural Language Processing (NLP) is important in this process because it helps us understand the structure, meaning, and context of the text. Using NLP techniques like tokenization, stop-word removal, and text normalization, we can turn raw text into a structured format that works well with machine learning algorithms.

This study is about creating a system to find cyberbullying The system checks what people write on the internet. It decides if the text is mean or not. It uses ways to understand the meaning of words. Then it uses algorithms like Naïve Bayes and Support Vector Machine to figure out if something is cyberbullying or not. The system was taught with examples of mean things people say online and things that are not mean. The main goal of this research is to make a system that can automatically find cyberbullying when people talk to each online. This can help websites like Facebook and Twitter monitor behavior. By getting better at finding mean things people say the system can help make the internet a nicer place for people to talk to each other help stop cyberbullying. and help social

media platforms, like Facebook and Twitter protect people from cyberbullying.

II. LITERATURE SURVEY

AUTHOR: Alqahtani and Ilyas

DESCRIPTION: Proposed a machine learning model to detect cyberbullying in social media text. Their approach used classifiers like Decision Trees, Random Forest and Support Vector Machine to get better prediction results. The ensemble model worked well and achieved higher performance compared to using individual classifiers like Decision Trees, Random Forest and Support Vector Machine. Cyberbullying detection, in media text benefited from combining these classifiers.

AUTHOR: Ates, Bostanci and Guzel

DESCRIPTION: I looked at how different machine learning algorithms do at finding cyberbullying. The people who did this research checked a lot of models to see how good they are. They used things like how accurate they're how precise they are if they can recall things and something called the F1-score. What they found out is that machine learning algorithms are really good at telling if something is bullying or not if you use the way to look at the words and stuff. Machine learning algorithms are very helpful for cyberbullying detection. The research on machine learning algorithms, for cyberbullying detection is important because it helps us understand how machine learning algorithms can be used for cyberbullying detection.

AUTHOR: Reddy et al.

DESCRIPTION: The person proposed a system to detect cyberbullying using machine learning techniques. They used natural language processing to get the text ready from media sites. The study used algorithms to find bullying messages. It showed that machine learning models are really good at finding language when people talk to each other online. The cyberbullying detection system is a way to stop people from being mean to each other, on the internet. The system uses cyberbullying detection to make social media a place.

III. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM: Reddy et al. created a system to find cyberbullying using machine learning. This system looks at what people write. Finds the bad messages. They showed that if you get the words ready and use the tools you can find the mean things people say online. Ates, Bostanci and Guzel. tried a few machine learning tools to see what works best for finding cyberbullying. They found out that if you pick the features and use the right methods you can get better at finding the bad stuff. These systems are pretty good, at finding cyberbullying in what people write. Most of them only look at what people write and do not look at pictures, sounds or videos. This is a problem because people use lots of types of media to talk to each other on social media like cyberbullying in pictures or videos which is also cyberbullying.

3.2 PROPOSED SYSTEM: The cyberbullying detection system that is being proposed is going to help us get around the problems that current systems have. This cyberbullying detection system can look at text and images and audio and video from media. It uses something called Natural Language Processing to get the text data ready. It uses something called TF-IDF to find the important parts of the text. The cyberbullying detection system uses machine learning to figure out if what people are saying is mean or not. It uses Naïve Bayes and Support Vector Machine to decide if the text is bullying or not. When it comes to images the cyberbullying detection system uses computer vision to find pictures and it uses Optical Character Recognition to read the words in memes. The cyberbullying detection system can also listen to audio. Turn it into text.. It can look at videos and take out the important parts. The cyberbullying detection system is going to be better at finding mean things because it can look at lots of different types of media. This is going to help social media platforms find things and make the internet a safer place for everyone.

IV. SYSTEM ARCHITECTURE

The cyberbullying detection system architecture is shown in Fig. 1. The system analyzes different types of social media data, such as text, images, audio, and videos. First, data is collected from social media platforms. The collected data is then cleaned using data preprocessing techniques. For audio data, speech-to-text methods are used to convert the audio into

textual form. Text embedded in images is extracted using Optical Character Recognition (OCR) techniques. In addition, image similarity methods are used to analyze visual patterns in images.The processed data is then sent to the cyberbullying detection module, where machine learning models analyze the data using a training dataset. Finally, the classification module determines whether the content is bullying or non-bullying based on the analysis performed by the detection module. This classification helps the system effectively identify cyberbullying content on social media platforms.

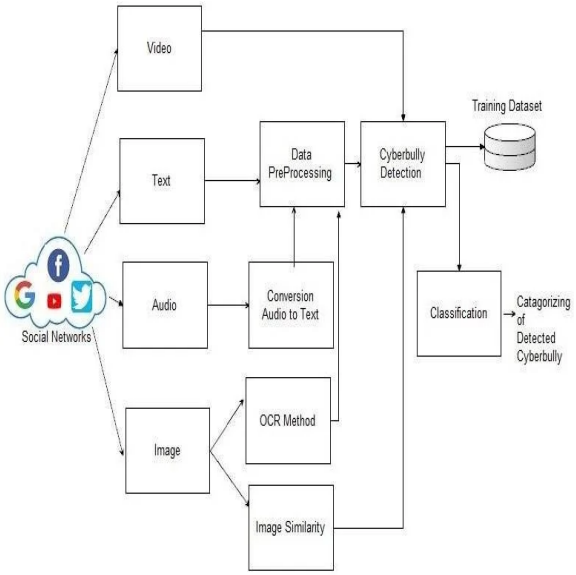

Fig. 1: System Architecture of Cyberbullying Detection Model

V. METHODOLOGY

The cyberbullying detection system is designed to identify harmful content on social media platforms. It analyzes multiple types of data, including text, images, audio, and video. The system applies machine learning and natural language processing techniques to process and analyze the collected data in order to detect cyberbullying behavior effectively.

1.Data Collection
The cyberbullying detection system is designed to identify harmful content on social media platforms. It analyzes multiple types of data, including text, images, audio, and video. The system applies machine learning and natural language processing techniques to process and analyze the collected data in order to detect cyberbullying behavior effectively.

2.Data Preprocessing
In this stage, the collected data is cleaned and prepared for analysis. Text preprocessing techniques such as tokenization, stop-word removal, and normalization are applied to the text data. Tokenization divides the text into smaller units, while stop-word removal eliminates common words such as “the” and “and” that do not contribute significant meaning. Normalization ensures that the text is converted into a consistent format. This process removes unnecessary words and prepares the text data for feature extraction and further analysis.

3.Audio to Text Conversion
For audio data, speech-to-text techniques are used to convert spoken words into textual form. This allows the system to analyze the content of audio data using natural language processing methods similar to those applied to written text.

4.OCR for Image Processing
For image data, the system uses Optical Character Recognition (OCR) to extract text from images or memes. The extracted text is then analyzed to identify abusive or harmful language.

5.Image Similarity Analysis
The system analyzes images to detect harmful or offensive visual content. It compares image features with known patterns associated with cyberbullying to identify potentially abusive images. This process helps the system recognize visual patterns that may indicate cyberbullying behavior.

6.Cyberbullying Detection
The collected data is first cleaned and relevant features are extracted. The processed data is then analyzed using machine learning algorithms to detect cyberbullying behavior. These algorithms are trained using a labeled dataset containing examples of bullying and non-bullying content, enabling the model to accurately identify harmful interactions online.

7.Classification
Finally, the system classifies the detected content into categories such as bullying or non-bullying. This

classification helps identify harmful content and enables social media platforms to take appropriate actions.

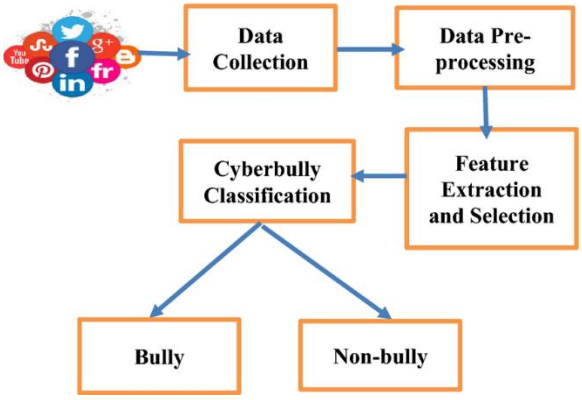

Fig. 2: Methodology of Cyberbullying Detection Model

VI. RESULT

The proposed cyberbullying detection model collects data from social media in various formats, including text, images, audio, and video. The data is pre-processed to remove irrelevant information and convert it into a suitable format, such as transforming audio into text and extracting text from images. Relevant features are then extracted to represent the data effectively. A machine learning algorithm is trained on this processed dataset to identify patterns associated with cyberbullying. The trained model analyses social media content and classifies it as either bullying or non-bullying, enabling automatic detection of harmful content and contributing to safer online platform.

VII. CONCLUSION

This study presents a system designed to detect cyberbullying on social media platforms. The proposed cyberbullying detection system uses machine learning techniques to analyze user-generated content from social media. It examines different types of data, including text, images, audio, and video. Natural language processing methods are used to understand and analyze written content, while additional techniques are applied to process visual and audio data. The system then applies machine learning algorithms to classify the content as cyberbullying or non-cyberbullying. By automatically identifying harmful content, the proposed system can assist social media platforms in monitoring online interactions and improving user safety, ultimately contributing to a safer online environment. The proposed system demonstrates how machine learning techniques can help address the problem of cyberbullying on social media platforms. It analyzes various types of user-generated content to identify harmful interactions between users. This approach supports the development of automated tools that monitor online activity and promote a safer and more respectful online community.

VIII. FUTURE WORK

Although the proposed system can detect cyberbullying across different types of media, there are several areas for improvement. Future work can focus on using advanced deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to improve detection accuracy. The system can also be enhanced by using larger and more diverse datasets from multiple social media platforms. In addition, implementing real-time monitoring and supporting multilingual detection would help the system identify cyberbullying more effectively and contribute to a safer online environment.

REFERENCES

[1] S. Salawu, Y. He, and J. Lumsden, “Approaches to Automated Detection of Cyberbullying: A Survey,” IEEE Transactions on Affective Computing 2020.

[2] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the Detection of Textual Cyberbullying,” in Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 2011.

[3] A. Reynolds, H. Kontosta this, and L. Edwards, “Using Machine Learning to Detect Cyberbullying,” in Proceedings of the IEEE International Conference on Machine Learning and Applications, 2011.

[4] P. Dadvar, D. Trieschnigg, and F. de Jong, “Improving Cyberbullying Detection with User Context,” in European Conference on Information Retrieval, 2013.

[5] H. Xu, M. Jun, X. Zhu, and A. Bellmore, “Learning from Bullying Traces in Social

Media," in Proceedings of NAACL Workshop on Language in social media, 2012.

[6] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," in Proceedings of NAACL-HLT, 2016.

[7] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," in Proceedings of ICWSM, 2017.

[8] N. Djuric et al., "Hate Speech Detection with Neural Networks," in Proceedings of the World Wide Web Conference, 2015.

[9] S. Agrawal and A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," in European Conference on Information Retrieval, 2018.

[10] E. Raisi and B. Huang, "Cyberbullying Detection with Weakly Supervised Machine Learning," in Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2018.

[11] H. Rosa et al., "Automatic Cyberbullying Detection: A Systematic Review," Computers in Human Behavior, vol. 93, pp. 333–345, 2019.

[12] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Computing Surveys, vol. 51, no. 4, 2018.

[13] B. Vidgen and T. Yasseri, "Detecting Hate Speech and Offensive Language on Social Media: A Systematic Review," Online Social Networks and Media, 2020.

[14] J. Badjatiya et al., "Deep Learning for Hate Speech Detection in Tweets," in Proceedings of the World Wide Web Conference Companion, 2017.

[15] T. Founta et al., "Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior," in Proceedings of the World Wide Web Conference, 2018.