



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 12, December 2025

Cyber-bullying Detection System using Deep Learning

Meghana K R, Kavya H V

Student, Dept. of MCA, PES Institute of Technology and Management, Shivamogga, Karnataka, India

Dept. of MCA, PES Institute of Technology and Management, Shivamogga, Karnataka, India

ABSTRACT: A Social media is one of the most widely used forms of communication. However, some individuals utilize these platforms maliciously; "cyber-bullying" is one example. The act of harassing or hurting someone online is known as cyber-bullying, and it is especially common among young people. Therefore, the goal of this research is to propose a model for detecting cyber-bullying that is based on deep learning algorithms.

I. INTRODUCTION

People online have more freedom to speak their minds than ever before. With social networks everywhere and information just a tap away, it's easy to say what you think. Plus, lots of folks use fake names or stay anonymous, which opens the door to all sorts of behavior good and bad. On the tech side, deep learning has changed the game. It lets machines learn from piles of unlabelled data, which means they can spot patterns or pick up on things people don't even notice. Researchers use these tools to sift through huge amounts of text, looking for things like hate speech or figuring out people's attitudes just by analyzing their words. But there's a dark side to all this. As social media grows, so does cyber-bullying. And it's not just one kind there are tons of ways that people harm one another on the internet. Victims experience actual suffering as well. Many people experience anxiety depression low self-esteem and more after being bullied in the worst cases, even thoughts of suicide

II. LITERATURE SURVEY

[1] P. Fortuna, and S. Nunes, "A survey on automatic detection of hate speech in text", ACM Computing Surveys (CSUR), 2018.

Fortuna and Nunes' (2018) survey provides a thorough overview of the literature on automatic hate speech recognition in text, highlighting the field's expanding importance given the growth of harmful online communication. The authors address the lack of a universally recognized definition of hate speech and offer a unified viewpoint to direct computational methods. They examine current datasets in several languages and emphasize the necessity for uniform tools and annotation standards.

[2] Perasso, (2020) "Cyber-bullying detection through machine learning: Can technology help to prevent internet bullying", Int J Manag Humanit, 2020.

The authors contend that the widespread use of social media and the fact that many victims or bystanders fail to report harassment quickly make cyber-bullying a major threat to teenagers' psychological wellbeing. Perasso contends that by examining vast amounts of internet communication data, automated technologies—especially machine learning—can be crucial in early diagnosis and prevention.

[3] S. Prashar, and S. Bhakar, "Real time cyber-bullying detection", Int J Eng Adv Technol, 2019.

Prashar and Bhakar emphasize how urgently automated methods for spotting cyber-bullying on social media sites are needed. The study highlights that because online communication is informal and dynamic common detection techniques like manual moderation and keyword-based filtering are inadequate. The authors use machine learning techniques to classify harmful or abusive content based on linguistic patterns keywords and contextual cues. Their approach places a strong emphasis on quick processing so that bullying behavior can be identified and reported immediately helping social media companies maintain safer and more effective online communication environments.

[4] Horea J. Yadav, D. Kumar, and D. Chauhan, "Cyber-bullying detection using pre-trained bert model", July, 2020.

The study draws attention to the shortcomings of earlier deep learning techniques and conventional machine learning which frequently rely significantly on human-generated features and have difficulty capturing the complex semantics of social media discourse. To address this the authors employ BERT (Bidirectional Encoder Representations from Transformers) a pre-trained contextual language model that has been improved on cyber-bullying datasets to differentiate between harmful and non-harmful information.

III. PROPOSED METHODOLOGY

The suggested model can accurately identify abusive harassing or threatening content even when it is concealed by slang emojis acronyms or indirect language by analyzing the semantic meaning and contextual connections within text. While borderline cases are referred to human evaluation a policy engine links severity levels to actions like warnings and temporary concealment. Before utilizing a trained deep learning model to identify it in real time the technology pre-processes social media text and transforms it into pertinent embedding's. Overall our suggested system offers a scalable intelligible and equitable method for cyber-bullying detection across various social media contexts by fusing cutting-edge NLP techniques adversarial resistance and ethical precautions.

Proposed Model Diagram

The core Cyber-bullying Detection System receives this information as well as the detection logs produced during processing. Administrators engage with the system concurrently by entering login credentials to gain access to the platform and keep an eye on activity. The system interacts with the database to store freshly generated data, including user information, detection logs, and forecast outcomes, and to retrieve saved data when needed. After processing the user text, the system sends it to the deep learning model, which assesses the content and provides a model output that indicates whether or not the text involves bullying behaviour. After that, the system receives this prediction and either displays it to the user or logs it for administrator review.

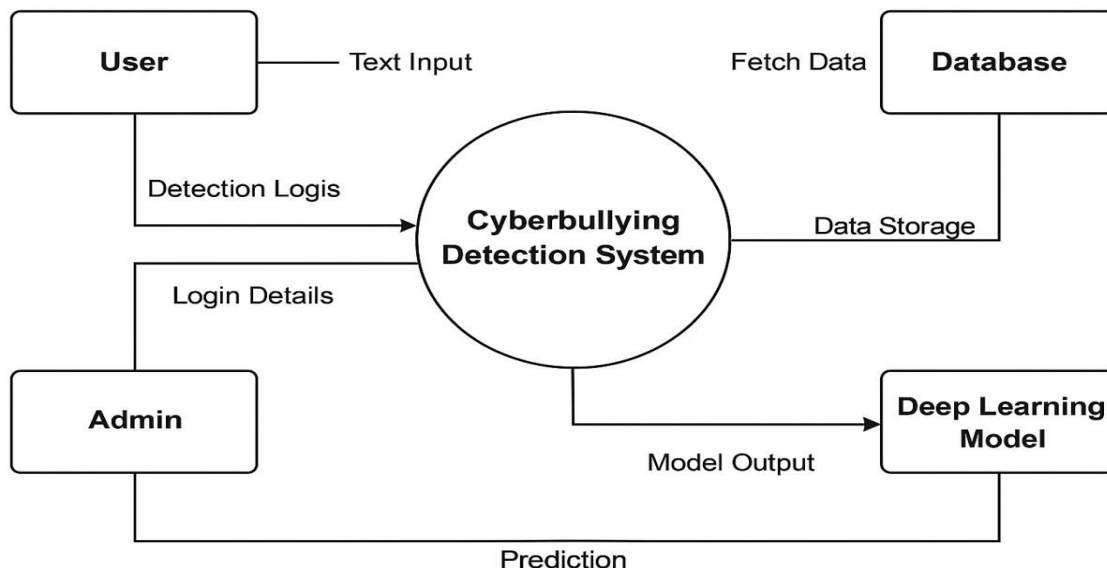


Figure 2.1.1: Block Diagram of Proposed System

IV. RESULTS

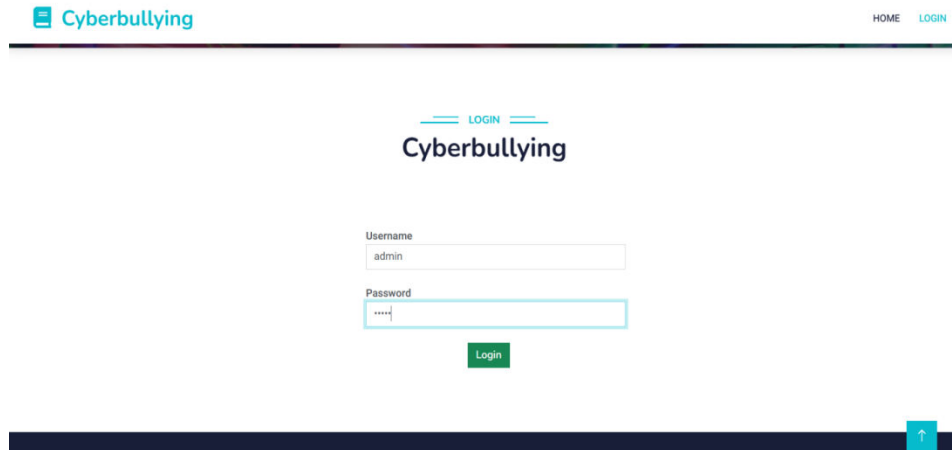


Fig: Login page

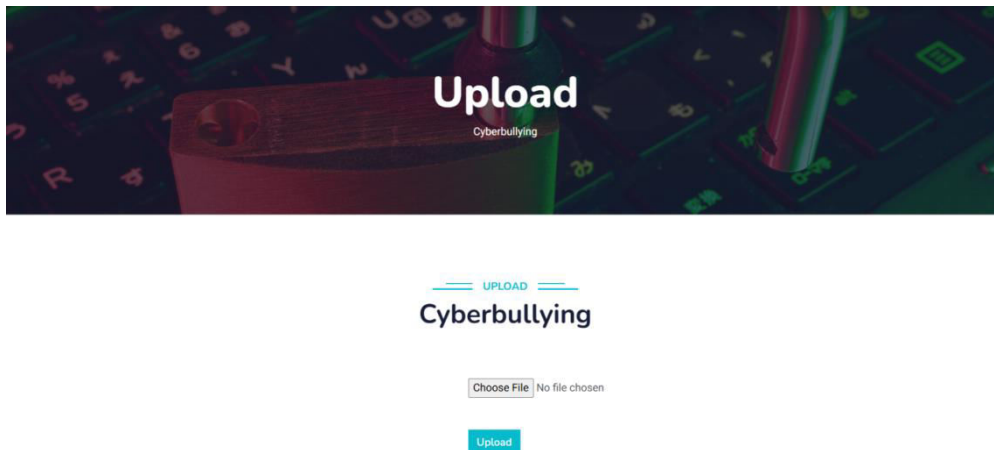
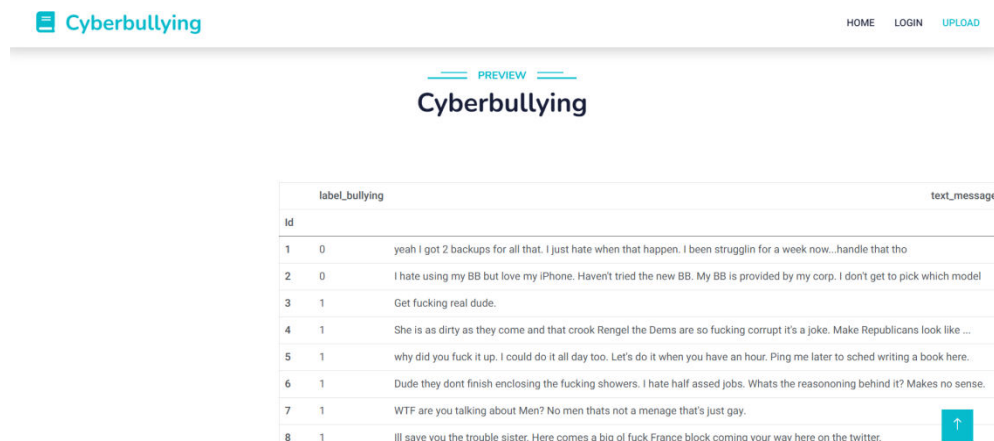


Fig: Upload Page



	label_bullying	text_message
id		
1	0	yeah I got 2 backups for all that. I just hate when that happen. I been strugglin for a week now...handle that tho
2	0	I hate using my BB but love my iPhone. Haven't tried the new BB. My BB is provided by my corp. I don't get to pick which model
3	1	Get fucking real dude.
4	1	She is as dirty as they come and that crook Rengel the Dems are so fucking corrupt it's a joke. Make Republicans look like ...
5	1	why did you fuck it up. I could do it all day too. Let's do it when you have an hour. Ping me later to sched writing a book here.
6	1	Dude they dont finish enclosing the fucking showers. I hate half assed jobs. Whats the reasoning behind it? Makes no sense.
7	1	WTF are you talking about Men? No men thats not a menage that's just gay.
8	1	Ill save you the trouble sister. Here comes a big ol fuck France block coming your way here on the twitter.

Fig: Check and Train Page

Cyberbullying

HOME LOGIN UPLOAD PREDICTION PERFORMANCE_ANALYSIS

PREDICTION

Enter your Text

hi how are you

submit

prediction is :
No_bullying

↑

Cyberbullying

HOME LOGIN UPLOAD PREDICTION PERFORMANCE_ANALYSIS

PREDICTION

Enter your Text

that sucks

submit

prediction is : bullying

↑

Fig: Prediction Bullying/no Bullying

Performance Analysis				
Cyberbullying				
PERFORMANCE ANALYSIS				
recall,F1 and Precision				
Recall f1 Precision				
0(Not_cyberbullying)	0.98	0.99	0.98	
1(Cyberbullying)	0.96	0.96	0.96	

Fig: Performance Analysis Page

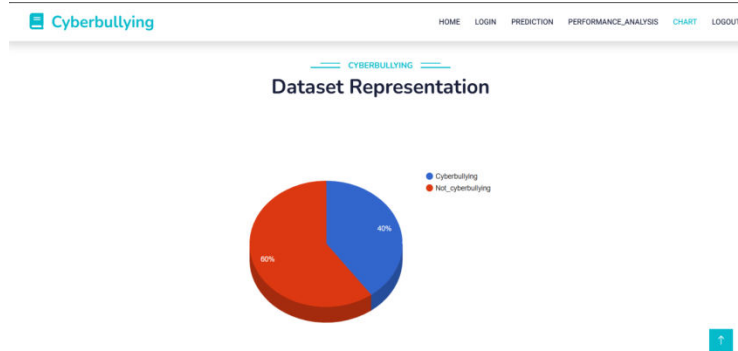


Fig: Dataset Representation Page

V. CONCLUSION

In conclusion, we proposed a semi-supervised strategy for recognizing cyber-bullying based on the five characteristics that the BERT model can use to recognize a post or message that contains cyber-bullying. When compared to traditional machine learning models the BERT model performed better trained across two cycles, attaining 91.90% accuracy while accounting for emotive features. The BERT model can generate more accurate findings if it is fed a large dataset. We may try to achieve even better results in the cyber-bullying detection approach if we include every component that we have recommended in this research piece. An application that can identify bullying traces and assist in finding and reporting such posts can be developed based on all the factors. By proactively detecting and reducing cyber-bullying behavior the study shows how deep learning technology can be used to create a safer online environment promoting responsible social media use and healthy digital communication.

Future Enhancement

To address the changing demands of online communication cyber-bullying detection systems can concentrate on enhancing flexibility scalability and ethical alignment. Integrating multimodal analysis processing text images audio and video at the same time is an essential tactic for identifying bullying that may manifest as memes voice notes or visual signs. The system will be more inclusive with greater support for multilingual and code-mixed languages particularly for diverse social media communities. Real-time active learning pipelines are another improvement that allows the system to constantly update its models with new vocabulary acronyms and emerging online aggressive patterns. Future research should focus on incorporating multilingual support which allows the model to detect bullying in regional languages mixed-language text and slang that frequently appears in social media interactions.

REFERENCES

- [1] M.E Kula., "Cyber-bullying: A Literature Review on Cross-Cultural Research in the Last Quarter", Handbook of Research on Digital Violence and Discrimination Studies, pp.610-630. 2022.
- [2] M. AGBAJE, and O. Afolabi, "Neural Network-Based Cyber-Bullying and Cyber-Aggression Detection Using Twitter Text", 2022.
- [3] A. Dewani, M. A. Memon, and S. Bhatti, "Cyber-bullying detection: advanced pre-processing techniques & deep learning architecture for Roman Urdu data", Journal of big data, Vol. 8, N. 1, pp. 1-20, 2021.
- [4] A.M Aldualaj, and A. Belghith, "Detecting Arabic Cyber-bullying Tweets Using Machine Learning", Machine Learning and Knowledge Extraction, Vol. 5, No. 1, pp. 29-42, 2023.
- [5] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyber-bullying detection solutions based on deep learning architectures", Multimedia Systems, 1-14, 2020.
- [6] M. Alotaibi, B. Alotaibi, and A. Razaque, "A multichannel deep learning framework for cyber-bullying detection on social media", Electronics, Vol. 10, No. 21, pp. 2664, 2021.
- [7] J. Yadav, D. Kumar, and D. Chauhan, "Cyber-bullying detection using pre-trained bert model", July, 2020
- [8] P. Fortuna, and S. Nunes, "A survey on automatic detection of hate speech in text", ACM Computing Surveys (CSUR), 2018



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details