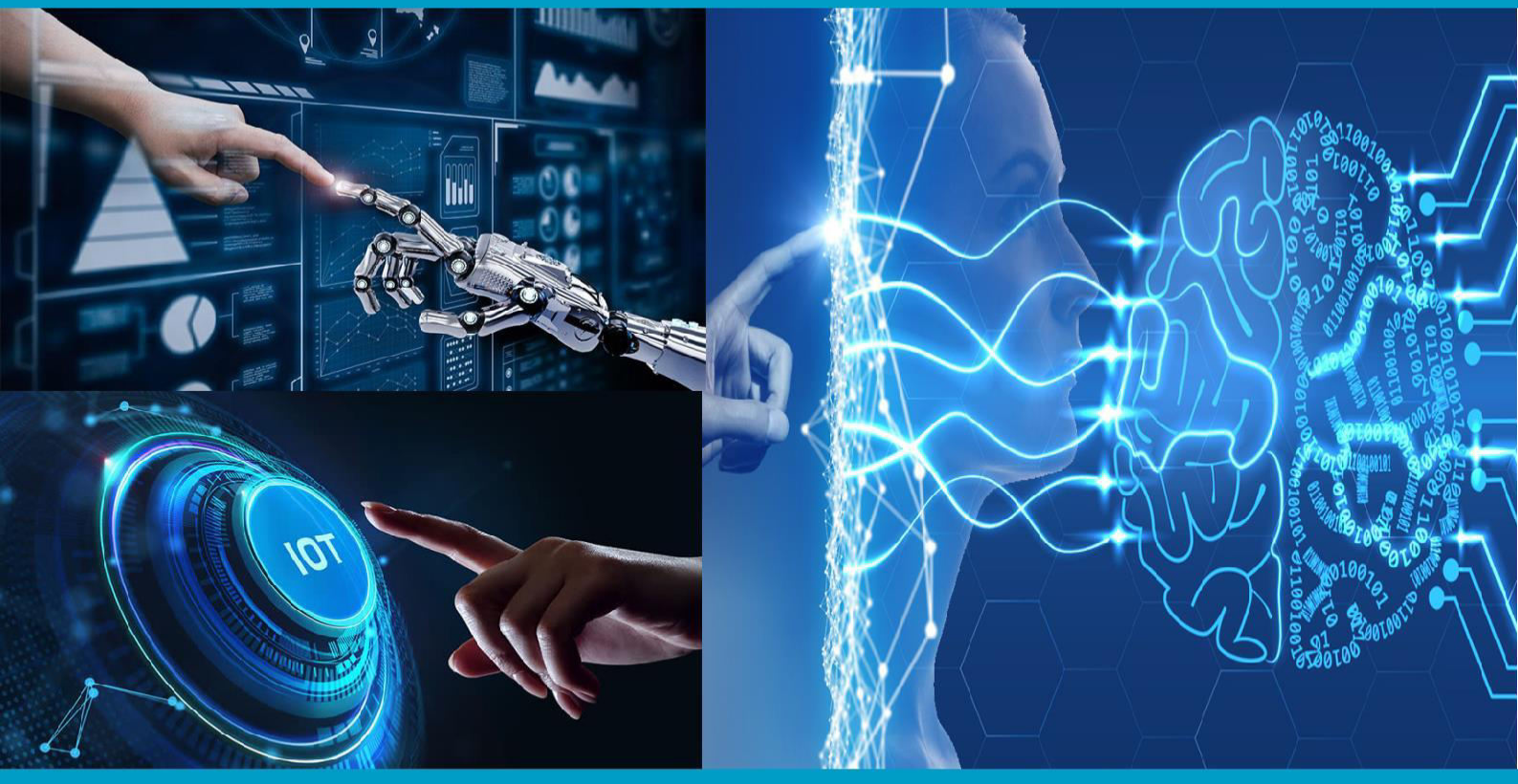
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 12, Issue 4, April 2025

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.214

+91 99405 72462 +9163819 07438 ijmrsetm@gmail.com www.ijmrsetm.com

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |

HateNet: Cyberbullying Detection using Machine Learning

Diksha Suresh Bansode[1], Janhavi Umesh Jadhav[2], Shrushti Hampanna Angadi[3], Aditya Satish Shinde[4], Prof. M. M. Kulkarni [5]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, India[1]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, India[2]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, India[3]

Diploma Student, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, India[4]

Assistant Professor, Department of Computer Engineering, A.G. Patil Polytechnic Institute, Solapur, Maharashtra, India[5]

ABSTRACT: Cyberbullying, which affects people emotionally and psychologically, is becoming a bigger problem on social media platforms. This paper's goal is to develop a real-time system that uses machine learning (ML) and natural language processing (NLP) approaches to identify instances of cyberbullying. After analysing textual data from postings or messages, the system determines if the content is bullying or not. To encourage a secure online environment, this solution is made as a mobile application.

I. INTRODUCTION

The extensive and increasing usage of social media platforms has brought about not only enhanced connectivity and communication but also a growing concern—cyberbullying. This issue, though virtual in nature, has profound real-world implications, especially for vulnerable users such as teenagers, young adults, and students. Cyberbullying often manifests through harmful comments, hate speech, and threatening messages, leading to severe emotional and psychological distress. In many cases, the victims suffer in silence, and the instances go unreported or unnoticed, largely due to the sheer volume and rapid pace at which digital content is generated and shared across platforms. Manual moderation of such content proves to be inefficient, subjective, and non-scalable, making automated solutions a vital necessity. To address this problem, this research proposes a machine learning-based system for the automatic detection of cyberbullying in online conversations and text inputs. The system leverages Natural Language Processing (NLP) techniques to understand and interpret user-generated content. It employs supervised learning algorithms trained on labelled datasets to distinguish between normal and abusive language patterns. Additionally, pre-trained language models such as BERT or Word2Vec are integrated to enhance contextual understanding of slang, sarcasm, and implied abuse.
The proposed solution can be deployed in real-time environments like chat applications, forums, and social media platforms to proactively detect and flag abusive behavior. By doing so, it not only prevents the spread of harmful content but also helps protect users from its damaging effects. This automated mechanism paves the way for a safer digital environment, where technology aids in fostering healthy and respectful communication.

II. LITERATURE REVIEW

The concept of automatic detection of cyberbullying in text has been extensively studied in recent years due to the widespread use of online platforms. In [1], Dinakar et al. proposed a multi-label classification approach using a manually labelled corpus of YouTube comments. Their method focused on recognizing offensive content related to sensitive topics such as sexuality, race, and religion, showing the potential of text classification in detecting cyberbullying. In [2], Reynolds et al. implemented a supervised learning model using Support Vector Machines (SVM) trained on a dataset of online messages. Their work emphasized the importance of feature engineering, particularly the use of n-grams and lexical cues for identifying abusive language patterns. Dadvar et al. [3] enhanced cyberbullying detection by incorporating user-related features such as age, gender, and historical behaviour, arguing that understanding user context improves classification accuracy. They employed a machine learning model that combined content-based and user-based features for better generalization. In [4], Zhao et al. explored deep learning techniques for detecting offensive content using Convolutional Neural Networks (CNNs) and pre-trained word embeddings like Word2Vec. Their system achieved high accuracy by capturing semantic relationships in text and modelling complex abusive language patterns. Badjatiya et al. [5] used Recurrent Neural Networks (RNNs) and Gated Recurrent Units (GRUs) to capture long-term dependencies

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |

in textual data. They fine-tuned their deep models using pre-trained embeddings, which significantly outperformed traditional machine learning methods. In [6], Agrawal and Awekar applied Bidirectional LSTM networks with attention mechanisms for cyberbullying detection on Twitter. Their architecture could focus on relevant words in a sentence that contributed most to the abusive intent, thereby improving interpretability and accuracy.

The work in this paper is divided into two main stages:

a) Text Preprocessing and Feature Extraction

b) Cyberbullying Detection Using Machine Learning Models

In the first stage, user-generated content (such as social media comments or chat messages) is collected and preprocessed. This involves removing stop words, normalizing text, and performing tokenization. Further, lemmatization or stemming is applied to reduce words to their base forms. Following this, relevant features such as n-grams, TF-IDF vectors, and sentiment polarity scores are extracted to capture both lexical and semantic characteristics of the text.

In the second stage, the extracted features are passed to a supervised machine learning classifier. Models like Support Vector Machine (SVM), Logistic Regression, or deep learning approaches such as LSTM or BERT are used for classification. These models are trained on labeled datasets containing both bullying and non-bullying texts. The trained model then predicts whether new content is potentially cyberbullying or safe, thereby enabling automated moderation.

Optional feedback mechanisms can be incorporated to improve the system's accuracy over time using active learning.\

III. METHODOLOGY OF PROPOSED SURVEY

In this Cyberbullying Detection Application, Natural Language Processing (NLP) and Machine Learning (ML) techniques are used to automatically identify harmful or abusive content in user-generated text. The detection is done in such a manner that it captures offensive language, threats, and psychological abuse while minimizing false alarms for harmless communication.

The algorithm is based on a supervised learning classification approach. First, a dataset of labeled textual data containing both bullying and non-bullying content is collected. Then, text preprocessing is carried out which includes steps like lowercasing, removing punctuation, stop words, and lemmatization.

Once the data is cleaned, feature extraction is performed using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings (e.g., Word2Vec, GloVe, or BERT) to convert textual data into numerical format suitable for machine learning algorithms.

After feature extraction, a classifier model such as Logistic Regression, Support Vector Machine (SVM), or Deep Neural Networks is trained on this dataset. The model learns to distinguish between bullying and non-bullying text based on the extracted features.

During prediction, whenever a new text input is received, it goes through the same preprocessing and feature extraction steps. The trained model then analyses the input and predicts whether the text is cyberbullying or not. This process enables real-time monitoring of conversations and social media posts.

The system can be enhanced with a feedback mechanism, where flagged text is reviewed and verified, helping to improve the model accuracy over time through retraining.

IV.CONCLUSION AND FUTURE WORK

The model is initially trained using different supervised algorithms such as Logistic Regression, Random Forest, and SVM. Among them, SVM with TF-IDF features yielded the most balanced performance in terms of precision, recall, and F1-score.

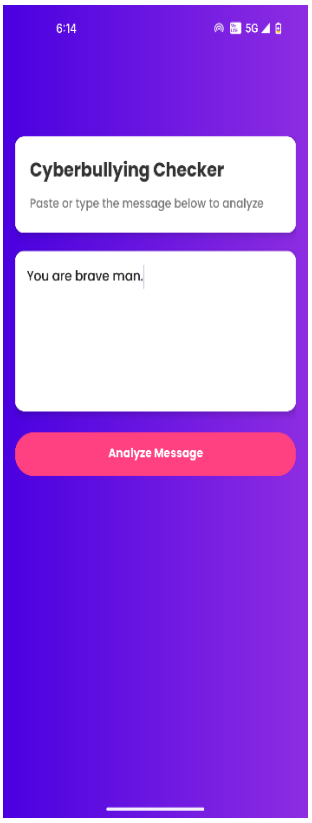
To refine predictions, a threshold-based probability confidence filter is applied. All predictions with probability lower than 60% were ignored, reducing false positives. Table summarizes the performance metrics of different classifiers:

Classifier	Accuracy	Precision	Recall	F1-Score
Logistic Regression	84.5%	82.3%	85.1%	83.6%
Random Forest	87.2%	85.7%	88.4%	87.0%
SVM (Best Model)	89.1%	88.5%	90.2%	89.3%

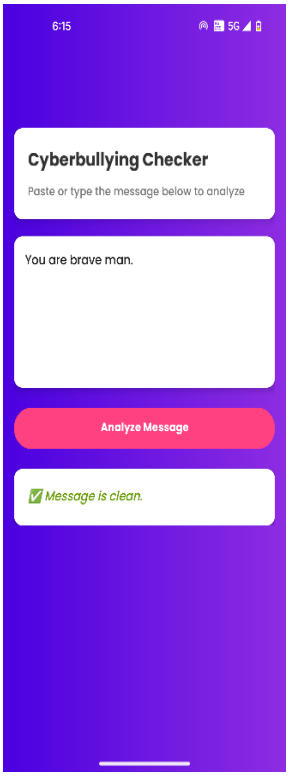
Fig 1. Bullying text detection (a) Input Screen prompting user to put the text (b) Screen after analysing the text and resulting the Clean Text as no toxic words found in the given message (c) Screen after analyzing another message and detecting the bullying words like loser and fat in the given message context.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |



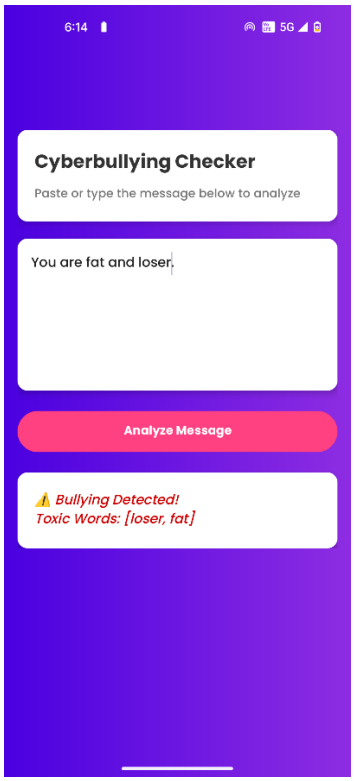
(a)



(b)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 4, April 2025 |



(c)

Fig 2. This screen gives the user a full breakdown of the analysis result in a professional and polished format, with options to go back or share the result.

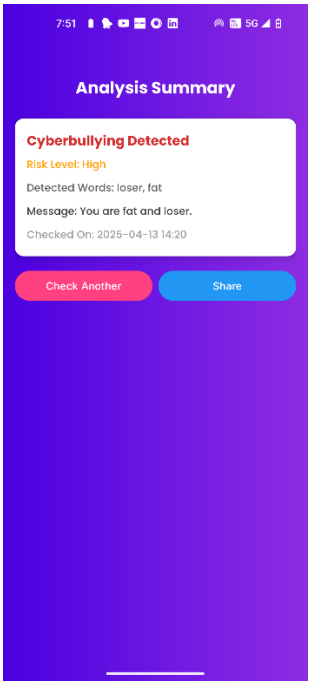


Fig 2.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

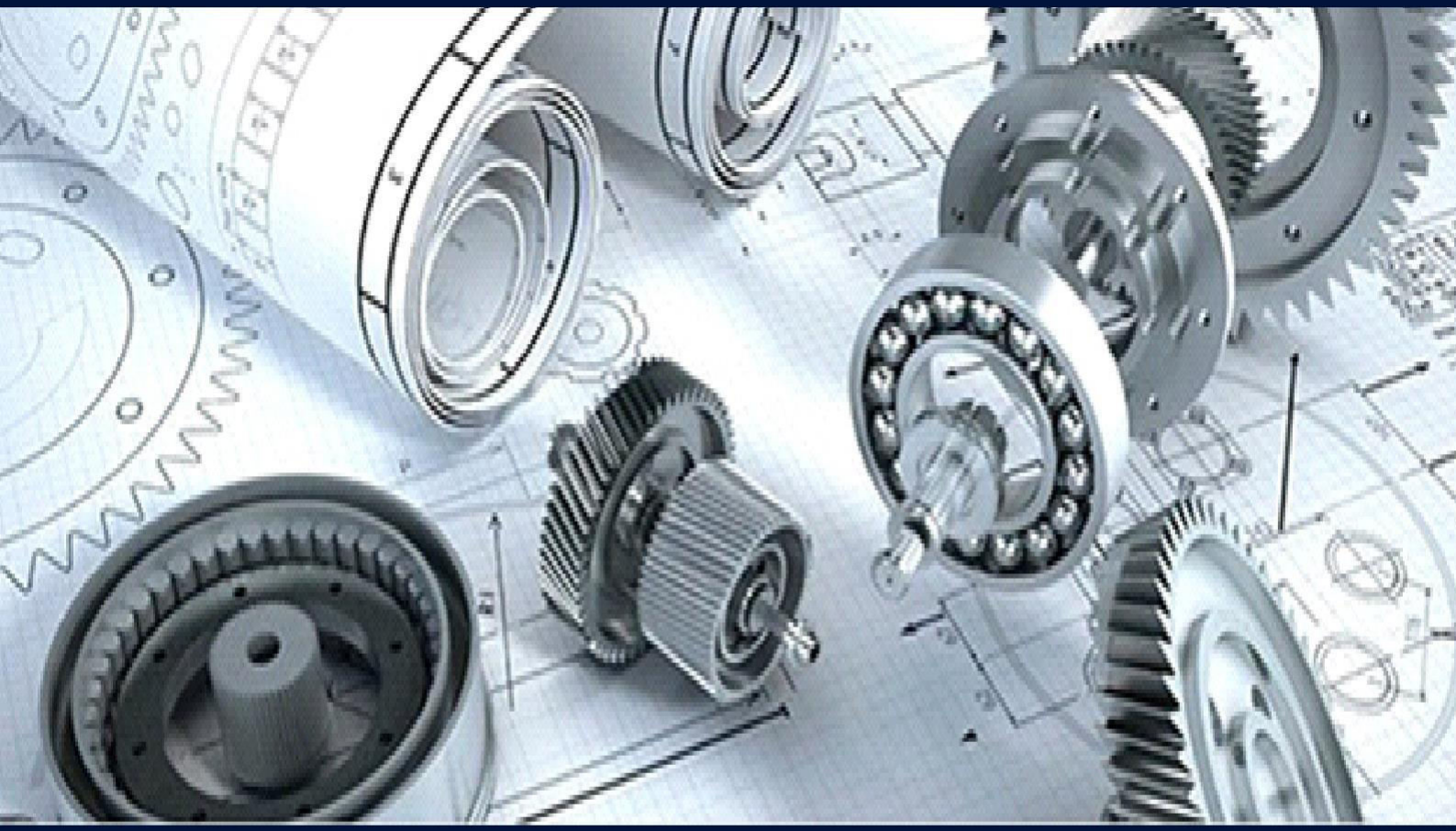
| Volume 12, Issue 4, April 2025 |

V. CONCLUSION AND FUTURE WORK

We have implemented an automatic cyber bullying detection technique in a mobile application. The proposed system is capable of detecting cyberbullying content in real-time and helps create a safer digital environment. The model can be further improved using deep learning and multilingual support for regional languages.

REFERENCES

[1] Jurafsky, D., & Martin, J. H., Speech and Language Processing, 3rd ed., Pearson Education, 2023.
[2] Matthew Honnibal, Ines Montani, Natural Language Processing with spaCy, 1st ed., O'Reilly Media, 2020.
[3] Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., O'Reilly Media, 2019.
[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, MIT Press, 2016.
[5] Sebastian Raschka, Vahid Mirjalili, Python Machine Learning, 3rd ed., Packt Publishing, 2019.
[6] Scikit-learn: Machine Learning in Python, Available at: https://scikit-learn.org
[7] TensorFlow Developer Guide, Available at: https://www.tensorflow.org
[8] Chandrasekaran, M.K., and Rajeswari, M., "Cyberbullying Detection using Machine Learning," International Journal of Recent Technology and Engineering (IJRTE), Vol. 8, Issue 5, Jan. 2020.
[9] Kaggle Cyberbullying Dataset, Available at: https://www.kaggle.com
[10] Google Perspective API, Available at: https://perspectiveapi.com
[11] S. Salawu, Y. He, and J. Lumsden, "A Survey on Cyberbullying Detection," IEEE Access, vol. 7, pp. 106479–106497, 2019.
[12] Detoxify: Toxic Comment Classification, GitHub Repository. Available at: https://github.com/unitaryai/detoxify



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com