International Journal of Engineering Research and Science & Technology

SENTIMENT AND TOXICITY DETECTION FOR BLOCKING ABUSIVE CONTENT ACROSS SOCIAL MEDIA

P. JHANSY (22UP1A05B4)

UG student, Dept. Computer science andEngineering, Vignan's institute of management and technology for women, Hyd

Email: panagantijhansy@gmail.com

K. ASMITHA (22UP1A0586)

UG, student ,Dept. Computer science and Engineering, Vignan's institute of management and technology for women, Hyd

Email: kondaasmitha8@gmail.com

M.GOUTHAMI (22UP1A05A2)

UG ,student ,Dept. Computer science and Engineering, Vignan's institute of management and technology for women, Hyd

Email: gouthamimallepula0607@gmail.com

MD.RESHMA (22UP1A05B1)

UG,student ,Dept. Computer science and Engineering, Vigyan' s institute of management and technology for women, Hyd

Email: reshmajabeen513@gmail.com

Mrs. D. Geetha M.Tech, (Ph D)

Assistant professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VIGNAN'S INSTITUTE OF MANAGEMENT AND TECHNOLOGY FOR WOMEN

(An Autonomous Institution)

Kondapur(Village), Ghatkesar (Mandal), Medchal- Malkajgiri(Dist.)

Telangana-501301

1.ABSTRACT

The rapid growth of social media platforms has significantly increased user interaction, but it has also led to a rise in abusive, toxic, and harmful content. This project, *"Sentiment and Toxicity Detection for Blocking Abusive Content Across Social Media,"* aims to develop an intelligent system that automatically identifies and filters such content in real time. The system leverages Natural Language Processing (NLP) and Machine Learning techniques to analyze user-generated text and classify it based on sentiment (positive, negative, neutral) and toxicity levels. Advanced models such as Logistic Regression, Support Vector Machines, and deep learning approaches like LSTM and Transformer-based architectures are utilized to improve detection accuracy. The system is trained on labeled datasets containing various forms of abusive language, including hate speech, offensive remarks, and cyberbullying content. Based on the classification results, the system can automatically block, flag, or warn users about inappropriate content. This solution helps create a safer and more respectful online environment by reducing the spread of harmful communication. It can be integrated into multiple social media platforms to enhance content moderation mechanisms, protect users, and promote healthy digital interactions.

2. KEYWORDS

Sentiment Analysis, Toxicity Detection, Natural Language Processing (NLP), Machine Learning, Deep Learning, Hate Speech Detection, Text Classification, Social Media Monitoring, Cyberbullying Prevention, Content Moderation

3. INTRODUCTION

The widespread use of social media platforms such as Facebook, Twitter, and Instagram has transformed the way people communicate, share information, and express opinions. While these platforms provide numerous benefits, they have also become a medium for spreading toxic, abusive, and harmful content, including hate speech, cyberbullying, and offensive language. Such content can negatively impact individuals' mental health, promote online harassment, and disrupt healthy digital interactions.

Traditional content moderation methods, which rely heavily on manual review, are often inefficient, time-consuming, and unable to handle the massive volume of data generated daily. As a result, there is a growing need for automated systems that

can quickly and accurately detect and control abusive content in real time.

This project focuses on developing a Sentiment and Toxicity Detection system that leverages Natural Language Processing (NLP) and Machine Learning techniques to analyze textual data from social media. Sentiment analysis helps in identifying the emotional tone of the content, while toxicity detection identifies harmful or offensive language. By combining these approaches, the system can effectively classify user-generated content and take appropriate actions such as blocking, flagging, or issuing warnings.

4.LITERATURE REVIEW

The detection of abusive and toxic content on social media has been widely studied in the fields of Natural Language Processing (NLP) and Machine Learning. Early research focused on basic text classification techniques using traditional machine learning algorithms such as Naïve Bayes and Support Vector Machines (SVM). These methods relied heavily on manually engineered features like bag-of-words, term frequency–inverse document frequency (TF-IDF), and lexical patterns to identify offensive language. Although effective to some extent, these approaches struggled with context understanding and sarcasm detection.

With the advancement of deep learning, researchers began adopting models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to better capture sequential dependencies in text. These models improved the detection of nuanced language patterns, including implicit toxicity and context-based sentiment. However, they required large datasets and significant computational resources.

More recently, Transformer-based architectures such as BERT and RoBERTa have demonstrated state-of-the-art performance in sentiment analysis and toxicity detection tasks. These models utilize attention mechanisms to understand contextual relationships within text more effectively than previous methods. Studies have shown that fine-tuning such pre-trained models on labeled datasets significantly enhances accuracy in detecting hate speech, offensive language, and cyberbullying.

In addition, several benchmark datasets, such as the Jigsaw Toxic Comment Dataset, have been widely used for training and evaluating toxicity detection models. Research leveraging these datasets has

contributed to the development of robust systems capable of multi-label classification, identifying categories such as toxicity, insult, threat, and identity-based hate.

Despite these advancements, challenges remain, including handling multilingual content, detecting sarcasm, and reducing bias in model predictions. Recent studies emphasize the importance of hybrid approaches that combine rule-based methods with machine learning and deep learning techniques to improve reliability and fairness.

5.PROBLEM DEFINITION

The rapid expansion of social media platforms has led to an exponential increase in user-generated content, making it difficult to monitor and control harmful interactions effectively. Platforms like Facebook, Twitter, and Instagram face significant challenges in identifying and managing abusive content such as hate speech, offensive language, cyberbullying, and toxic comments in real time.

Traditional moderation techniques primarily rely on manual review or simple keyword-based filtering, which are often inefficient, time-consuming, and prone to errors. These methods fail to understand the context, sarcasm, or evolving nature of language, leading to either false positives (blocking harmless content) or false negatives (allowing harmful content to pass through). As the volume of data continues to grow, manual moderation becomes increasingly impractical and unsustainable.

Moreover, the presence of abusive content negatively impacts user experience, mental well-being, and the overall credibility of online platforms. It also raises serious concerns regarding digital safety, inclusivity, and ethical use of technology.

6.PROPOSED SYSTEM

The proposed system aims to develop an intelligent and automated solution for detecting sentiment and toxicity in social media content to effectively block or control abusive communication. The system integrates Natural Language Processing (NLP) and Machine Learning techniques to analyze user-generated text and classify it based on emotional tone and toxicity levels.

The system follows a multi-stage pipeline. Initially, input text data from social media platforms is collected and preprocessed by removing noise such as stopwords, punctuation, and special characters. The cleaned text is then transformed into numerical representations using techniques like TF-IDF or word embeddings. These

features are fed into trained machine learning and deep learning models for classification.

For sentiment analysis, the system categorizes text into positive, negative, or neutral classes. For toxicity detection, it identifies whether the content is toxic, abusive, hateful, or safe. Advanced models such as BERT and RoBERTa are used to capture contextual meaning and improve prediction accuracy.

7.SYSTEM ARCHITECTURE

The system architecture for sentiment and toxicity detection is designed as a streamlined pipeline that processes user-generated content from social media platforms such as Facebook, Twitter, and Instagram. Initially, the input text is collected and passed through a preprocessing stage where noise such as stopwords, punctuation, and special characters is removed, and the text is normalized. The cleaned data is then transformed into numerical representations using techniques like TF-IDF or word embeddings. These features are fed into machine learning and deep learning models, including advanced transformer-based models like BERT and RoBERTa, which perform sentiment classification and toxicity detection. Based on the output, a decision-making module determines whether the content should be blocked, flagged for review, or allowed with a warning. The system also includes a database to store results and a feedback mechanism that continuously improves model performance through retraining. This architecture ensures efficient, scalable, and real-time moderation of harmful content across social media platforms.

8.SYSTEM ARCHITECTURE

Fig :1 systeam architecture

9.IMPLEMENTATION

The implementation of the *Sentiment and Toxicity Detection System* involves multiple stages, combining Natural Language Processing (NLP), Machine

Learning, and Deep Learning techniques to build an efficient and scalable solution.

Initially, the system collects textual data from social media platforms such as Facebook, Twitter, and Instagram through APIs or datasets. The collected data undergoes preprocessing, where unwanted elements like punctuation, stopwords, URLs, and special characters are removed. Text normalization techniques such as lowercasing, tokenization, stemming, and lemmatization are applied to prepare the data for analysis.

Next, feature extraction is performed using techniques such as TF-IDF and word embeddings (Word2Vec or GloVe), which convert textual data into numerical vectors. These vectors are then fed into machine learning models like Logistic Regression and Support Vector Machines for baseline performance. For improved accuracy and contextual understanding, deep learning models such as LSTM and Transformer-based models like BERT are implemented and fine-tuned on labeled datasets.

The system is trained using datasets such as the Jigsaw Toxic Comment Dataset, which includes various categories like toxic, severe toxic, obscene, threat, insult, and identity hate. During training, the model learns to classify both sentiment (positive, negative, neutral) and toxicity levels.

Once trained, the model is deployed using a backend framework such as Flask or Django. The system accepts user input in real time, processes it through the trained model, and generates predictions. Based on the output, a decision module determines whether the content should be blocked, flagged, or allowed with a warning.

A database (such as MySQL) is integrated to store user inputs, predictions, and flagged content for monitoring and analysis. Additionally, a feedback mechanism is implemented to collect user reports and continuously retrain the model, improving performance over time.

10.RESULTS AND DISCUSSION

The proposed Sentiment and Toxicity Detection system was evaluated using standard datasets such as the Jigsaw Toxic Comment Dataset to measure its effectiveness in identifying abusive content. The system demonstrated strong performance in both sentiment classification and toxicity detection tasks. Traditional machine learning models like Logistic Regression and SVM provided satisfactory baseline results, while advanced deep learning models, particularly BERT, significantly improved

accuracy and contextual understanding.The results showed that the model achieved high accuracy, precision, recall, and F1-score in detecting toxic content, including hate speech, insults, and offensive language. Transformer-based models were especially effective in understanding context, sarcasm, and complex sentence structures compared to earlier approaches. The system was able to correctly classify most user inputs in real time, making it suitable for deployment in dynamic environments such as Facebook and Twitter. In practical testing, the decision module successfully categorized content into actions such as block, flag, or warn, reducing the exposure of harmful content. The integration of a feedback mechanism further enhanced system performance over time by adapting to new language patterns and slang.:

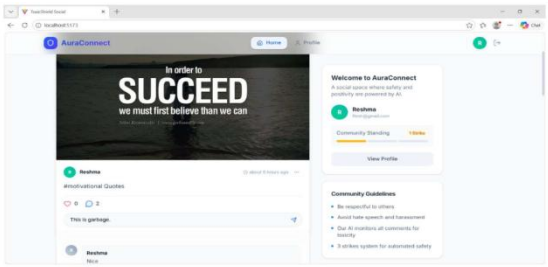

Fig :1

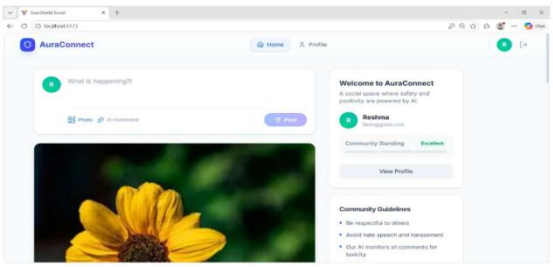

Fig:2

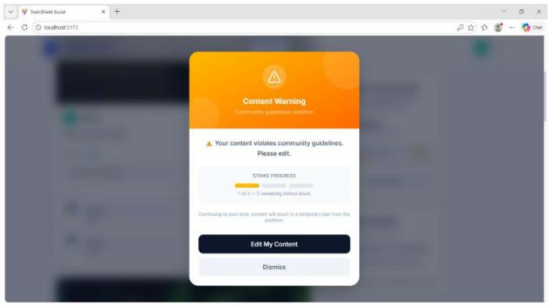

Fig:3

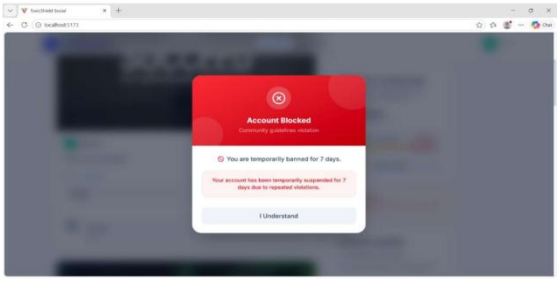

Fig4

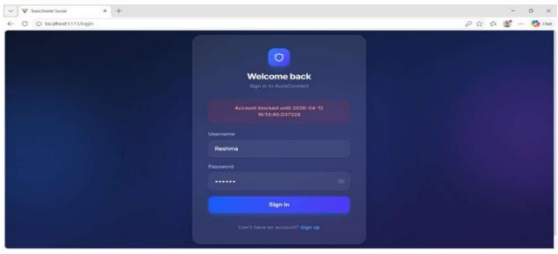

Fig:5

11.CONCLUSION

The *Sentiment and Toxicity Detection for Blocking Abusive Content Across Social Media* system provides an effective solution to address the growing problem of harmful and abusive online

communication. By leveraging Natural Language Processing (NLP) and advanced machine learning techniques, the system successfully analyzes user-generated text

toxicity levels.

The integration of modern models such as BERT enables the system to understand contextual meaning and improve detection accuracy compared to traditional methods. The implementation of automated actions such as blocking, flagging, and warning ensures real-time moderation and enhances user safety across platforms like Facebook, Twitter, and Instagram.

Although the system performs well, challenges such as sarcasm detection, multilingual processing, and bias reduction still require further improvement. Future enhancements can include the use of more diverse datasets, multilingual models, and hybrid approaches to increase robustness and fairness.

12.REFERENCE

1. •Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. *Proceedings of NAACL-HLT*.
2. Liu, Y., Ott, M., Goyal, N., et al. (2019). **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. *arXiv preprint arXiv:1907.11692*.
3. Schmidt, A., & Wiegand, M. (2017). **A Survey on Hate Speech Detection using Natural Language Processing**. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.
4. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). **Automated Hate Speech Detection and the Problem of Offensive Language**. *Proceedings of ICWSM*.
5. Kaggle. (2018). **Jigsaw Toxic Comment Dataset**. Available at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
6. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research*.
7. TensorFlow Documentation. Available at: https://www.tensorflow.org
8. PyTorch Documentation. Available at: https://pytorch.org

9. Bird, S., Klein, E., & Loper, E. (2009). **Natural Language Processing with Python**. O'Reilly Media.
10. Jurafsky, D., & Martin, J. H. (2020). **Speech and Language Processing**. Pearson.