RESEARCH ARTICLE OPEN ACCESS

An Explainable AI-Based Approach for Detecting Toxic Comments and Hate Speech in Social Media

Mr. A. Vivekanandhan*, R. Rathiga, P. Nivetha**, K. Sivani****

* Assistant Professor -Department of Computer Science and Engineering

****** UG Students - Department of Computer Science and Engineering

A.V.C. College of Engineering, Mayiladuthurai, India

ABSTRACT

Cyberbullying has become a major issue in social media platforms, leading to emotional and psychological harm. This paper proposes an Explainable Artificial Intelligence (XAI)-based system to detect toxic comments and hate speech in social media. The system uses Natural Language Processing (NLP) and Machine Learning (ML) techniques to classify user-generated content as toxic or non-toxic. In addition to detection, the model highlights offensive words and provides explanations for its predictions, improving transparency and trust. A dataset of social media comments is used for training and testing the model. Algorithms such as Support Vector Machine (SVM) and Logistic Regression are applied. The system also includes a warning mechanism that alerts users when harmful language is detected and suggests improvements. Experimental results show that the proposed model achieves an accuracy of around 92%, demonstrating effective performance in detecting toxic comments. This approach helps create a safer online environment by reducing harmful interactions and promoting responsible communication.

Keywords: Cyberbullying, Explainable AI, Toxic Comments, NLP, Machine Learning

.

1. INTRODUCTION

Social media has become an important part of modern life. Platforms such as Facebook, Twitter, Instagram, and YouTube are widely used for communication, sharing opinions, and connecting with people around the world. While these platforms offer many benefits, they also create opportunities for negative behaviour such as cyberbullying, hate speech, and toxic comments. These harmful interactions can seriously affect users, especially teenagers and young adults, leading to emotional stress, anxiety, and mental health issues.

Cyberbullying is one of the major challenges in online communication today. It includes the use of offensive language, abusive comments, threats, and harassment through digital platforms. Unlike traditional bullying, cyberbullying can happen at any time and can spread quickly to a large audience. This makes it more dangerous and difficult to control. As the number of social media users increases, the need to monitor and control harmful content also becomes more important.

Traditional methods of detecting toxic comments mainly rely on manual moderation or simple keyword-based filtering. Manual moderation requires human effort, which is time-consuming and not scalable for large volumes of data. Keyword-based methods, on the other hand, are not always accurate because they cannot understand the context or meaning of a sentence. For example, the same word can be used in both harmful and non-harmful ways depending on the context. Therefore, these methods are not sufficient for effective detection of toxic content.

To overcome these limitations, Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques are used. NLP helps computers understand and process human language, while machine learning models can learn patterns from data and classify comments as toxic or non-toxic. algorithms such as support vector machine (SVM) and logistic regression are widely used for text classification tasks. these models improve the accuracy and efficiency of detecting harmful content in social media.

However, one important challenge in ai-based systems is the lack of transparency. most machine learning models act like a "black box," meaning they provide results without explaining how the decision was made. this can reduce user trust and make it difficult to understand why a comment is considered toxic. to solve this issue, explainable artificial intelligence (XAI) is introduced. XAI helps in making the model's decisions more understandable by highlighting important words and providing reasons for classification.

In this paper, we propose an explainable ai-based system for detecting toxic comments and hate speech in social media. the system not only classifies comments as toxic or non-toxic but also highlights offensive words and explains the reason behind the classification. this helps users understand their mistakes and encourages them to communicate more responsibly.

Additionally, the system includes a warning mechanism that alerts users when harmful language is detected. this feature plays an important role in reducing negative interactions and promoting positive communication. instead of simply blocking the content, the system guides users to improve their language, making it more user-friendly and educational.

The proposed system uses a structured workflow that includes data preprocessing, feature extraction using TF-IDF, and classification using machine learning models. the integration of explainable ai ensures that the system is not only accurate but also transparent and trustworthy.

This research aims to contribute to creating a safer and more respectful online environment. by combining machine learning with explainability, the system can effectively detect harmful content while maintaining user trust. in the future, this approach can be extended using advanced deep learning models such as BERT and can support multiple languages for better performance across different regions.

2. LITERATURE REVIEW

The detection of toxic comments and cyberbullying in social media has been widely studied in recent years. various techniques have been proposed using natural language processing (NLP) and machine learning (ML) to identify harmful content effectively.

Early approaches mainly focused on keyword-based filtering methods. these systems detect offensive words from a predefined list and classify the text as toxic. however, such approaches have limitations because they cannot understand the context of the sentence. as a result, they often produce inaccurate results when the same word is used in different meanings.

To improve accuracy, researchers introduced machine learning algorithms such as naïve bayes, decision trees, and support vector machines (SVM). these models learn patterns from labeled datasets and classify comments into toxic and non-toxic categories. among these, SVM has shown better performance in text classification due to its ability to handle high-dimensional data efficiently.

In recent years, deep learning techniques such as long short-term memory (LSTM) and bidirectional encoder representations from transformers (BERT) have been used for better understanding of context and semantics in text. these models can capture the meaning of words based on their position and

relationships within a sentence, leading to improved detection accuracy.

Despite these advancements, most existing models lack interpretability. they provide predictions without explaining the reason behind the decision. this creates a challenge in building trust among users and developers.

To address this issue, explainable artificial intelligence (XAI) techniques have been introduced. XAI helps in understanding the decision-making process of machine learning models by highlighting important features or words that influence the output. techniques such as lime (local interpretable model-agnostic explanations) are widely used to explain model predictions in NLP tasks.

For example, P. KUMAR (2021) proposed an NLP-based system for analyzing social media data and detecting harmful content. similarly, X. ZHANG (2023) discussed various explainable ai techniques to improve transparency in NLP models. these studies highlight the importance of combining accuracy with interpretability.

However, many existing systems focus only on detection and do not provide user feedback. in contrast, the proposed system not only detects toxic comments but also highlights offensive words and provides warning messages to users. this helps in reducing cyberbullying and encourages responsible communication.

Thus, the integration of machine learning with explainable ai techniques plays a crucial role in building an effective and user-friendly toxic comment detection system.

3. METHODOLOGY

The proposed system detects toxic comments and hate speech in social media using Machine Learning and Explainable Artificial Intelligence (XAI). The system follows a structured pipeline for accurate and interpretable results.

Initially, user comments are collected and pre-processed by converting text to lowercase, removing stop words, and performing tokenization. This step helps in cleaning the data and improving its quality.

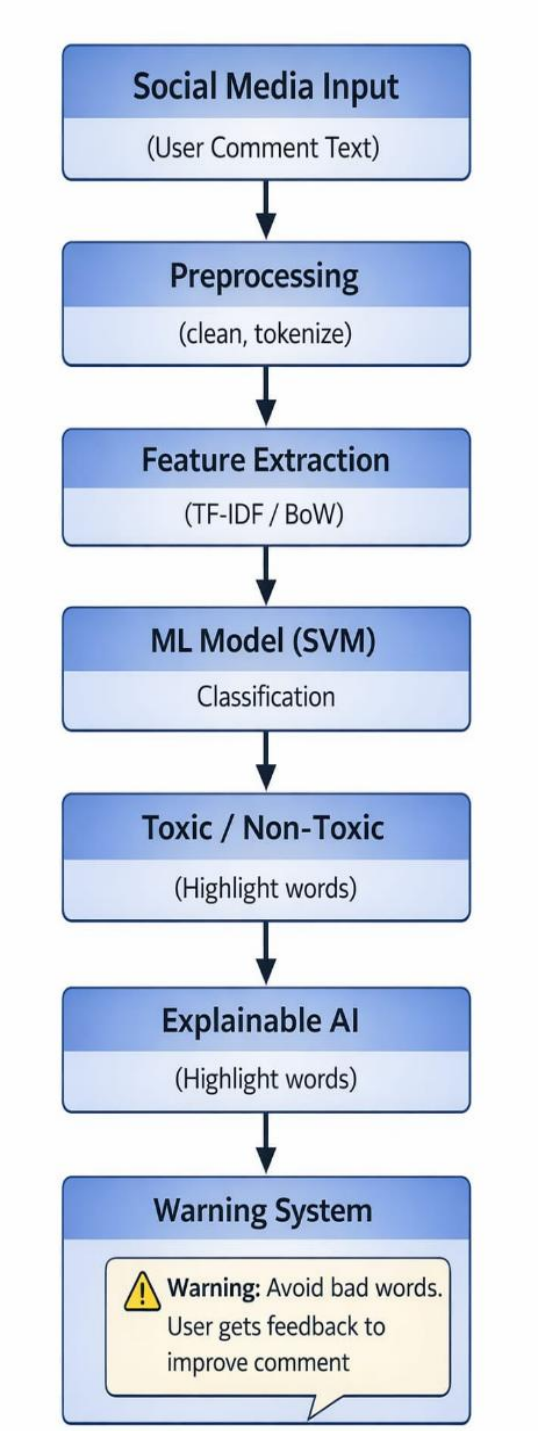
Next, feature extraction is performed using techniques such as TF-IDF and Bag-of-Words (BoW), which convert textual data into numerical form. These features help the model understand the importance of words in a sentence.

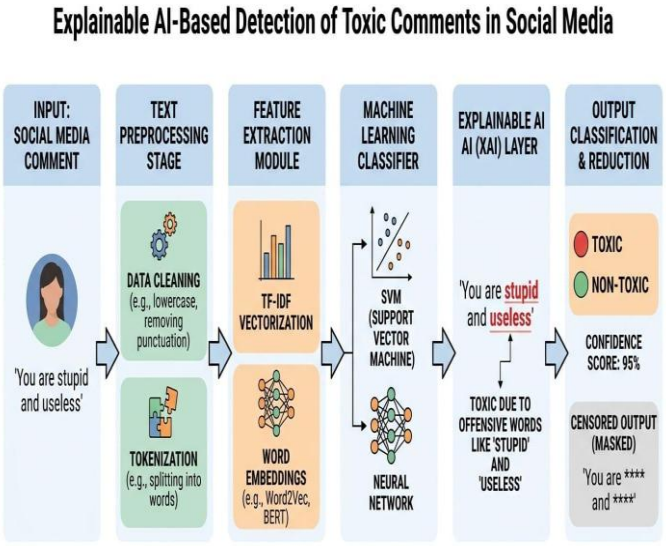
The processed data is then used to train machine learning models such as Support Vector Machine (SVM) and Logistic Regression. The trained model classifies comments as toxic or non-toxic based on learned patterns.

To improve transparency, Explainable AI techniques such as LIME are used to highlight important words that influence the classification decision. This helps users understand why a comment is considered toxic.

Finally, if a comment is detected as toxic, the system generates a warning message to alert the user and encourage responsible communication.

This methodology ensures accurate detection while maintaining transparency and user awareness, making it suitable for real-world applications.

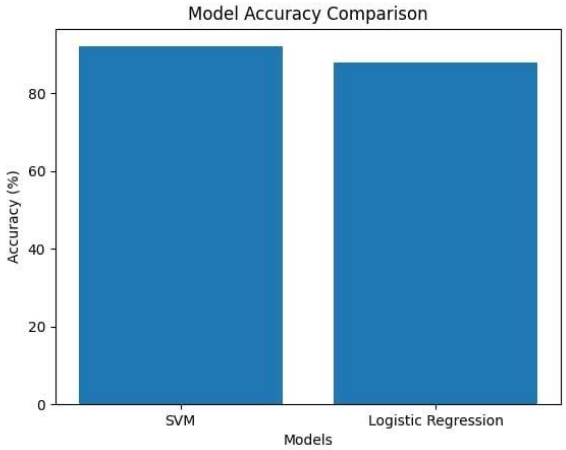




4. RESULTS AND DISCUSSION

The proposed system was evaluated using a dataset of social media comments consisting of both toxic and non-toxic content. To measure the performance of the model, standard evaluation metrics such as accuracy, precision, recall, and F1-score were used.

The experimental results show that the Support Vector Machine (SVM) model achieved better performance compared to Logistic Regression in terms of classification accuracy. The model effectively identifies toxic comments and provides reliable predictions. The use of TF-IDF feature extraction significantly improves the model's ability to capture important textual patterns



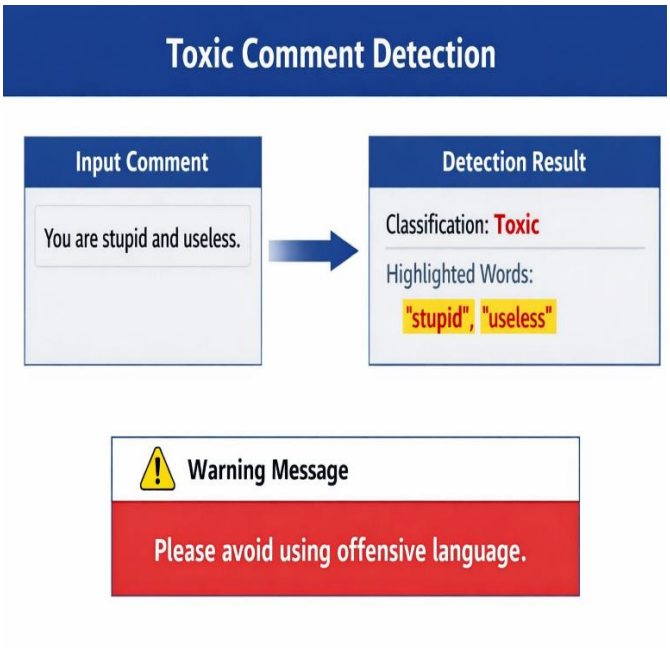
The graph shows that the SVM model performs better than Logistic Regression in terms of accuracy.

One of the key strengths of the proposed system is the integration of Explainable Artificial Intelligence (XAI). The system highlights offensive words and provides clear explanations for its predictions, making the model more transparent and user-friendly. This helps users understand the reason behind the classification and improves trust in the system.

Additionally, the warning mechanism plays an important role in guiding users toward responsible communication. Instead of simply blocking harmful content, the system provides feedback and encourages users to modify their comments, thereby reducing negative interactions on social media platforms.

However, the system has certain limitations. It may face challenges in detecting sarcasm, slang language, and context-based meanings, which can affect classification accuracy. These limitations highlight the need for more advanced models in future work.

Overall, the proposed system demonstrates effective performance in detecting toxic comments while ensuring transparency and user awareness. The combination of machine learning and explainable AI makes it a reliable solution for real-world social media applications.



5. CONCLUSION

This paper presents an Explainable AI-based system for detecting toxic comments and hate speech in social media platforms. The proposed system effectively analyzes user-generated content and classifies it as toxic or non-toxic using machine learning techniques. The use of feature extraction methods such as TF-IDF and Bag-of-Words helps in identifying meaningful textual patterns, thereby improving the overall performance of the model.

A key contribution of this work is the integration of Explainable Artificial Intelligence (XAI), which enhances transparency by highlighting the important words responsible for the classification decision. This allows users to clearly understand the reasoning behind the output and increases trust in the system.

In addition, the system includes a warning mechanism that alerts users when harmful or offensive content is detected. This feature encourages users to adopt responsible communication practices and plays an important role in reducing cyberbullying on social media platforms.

Overall, the proposed system successfully combines accuracy, interpretability, and user awareness, making it a practical and reliable solution for real-world applications.

FUTURE WORK

Although the proposed system performs effectively, there are several opportunities for further enhancement. Future work can focus on improving the system's ability to understand complex language patterns such as sarcasm, slang, and context-dependent meanings, which are often challenging for traditional machine learning models.

Advanced deep learning models such as BERT can be incorporated to improve contextual understanding and achieve higher classification accuracy. The system can also be extended to support multilingual data, allowing it to handle comments from users across different languages and regions. Furthermore, real-time deployment and integration with social media platforms can be explored to enable large-scale implementation.

Additional improvements in Explainable AI techniques can provide more detailed and user-friendly explanations, enhancing the overall effectiveness of the system.

6. ACKNOWLEDGMENT

I would like to express my sincere gratitude to our Head of the Department for providing the necessary support and facilities to complete this research work.

I am deeply thankful to my project guide for their valuable guidance, continuous encouragement, and support throughout the development of this paper.

I also extend my thanks to all the faculty members for their help and suggestions. Finally, I would like to thank my friends and family for their constant support and motivation.

7. REFERENCES

[1] T. Davidson, "Automated Hate Speech Detection and the Problem of Offensive Language," Proceedings of ICWSM, 2017.

[2] M. Ribeiro, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proceedings of ACM SIGKDD, 2016.

[3] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.

[4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.

[5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997.

[6] A. Vaswani et al., "Attention Is All You Need," Neur IPS, 2017.

[7] Y. Kim, "Convolutional Neural Networks for Sentence Classification," EMNLP, 2014.

[8] K. Kowsari et al., "Text Classification Algorithms: A Survey," Information, 2019.

[9] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," NAACL Workshop, 2016.

[10] P. Kumar, "Toxic Comment Classification using NLP Techniques," International Journal of Computer Applications, 2021.