International Journal of Advanced Research and Review

www.ijarr.org

Social Media Toxic Comment Classifier

[1]Tata Lavanya, [2]Kayala Chaitanya Swamy, [3]Pothini Tharun Sai, [4]Dr. M. Ratna Raju

[1,2,3]U. G Student, Dept COMPUTER SCIENCE AND ENGINEERING, St.Ann's College Of Engineering and Technology, Nayunipalli (V), Vetapalem (M), Chirala, Bapatla Dist, Andhra Pradesh – 523187, India

[4]Associate professor, COMPUTER SCIENCE AND ENGINEERING, St.Ann's College Of Engineering and Technology, Nayunipalli (V), Vetapalem (M), Chirala, Bapatla Dist, Andhra Pradesh – 523187, India

ABSTRACT

The rapid growth of social media platforms has enhanced communication but also increased the spread of toxic and abusive content. This project develops a social media toxic comment classifier using natural language processing (NLP) and machine learning to automatically detect and categorize comments as toxic, sever toxic, obscene, threat, insult, or identity hate. Using the jigsaw Toxic Comment Classification Challenge dataset, the system performs data preprocessing, tokenization, feature extraction, and classification through algorithms like Logistic Regression, Naïve Bayes, and deep learning models. The classifier achieves high accuracy and can be integrated into online platforms to promote healthier digital interactions, helping reduce online toxicity and foster safer, more inclusive communities.

KEYWORDS *Toxic Comment Classification, Hate Speech Detection, Natural Language Processing (NLP), Machine learning, Text Preprocessing, tokenization.*

INTRODUCTION

Many conversations take place on social media, and the anonymity of these platforms has allowed many people to freely share their thoughts. However, this freedom can occasionally

be abused by those who strongly disagree with a viewpoint. With the ongoing risk of harassment or hurtful remarks on the internet, sharing things that are important to you will become challenging. People will eventually cease asking for other people's opinions on their ideas and stop discussing them online as a result of this. Unfortunately, social media sites constantly deal with these problems and struggle to spot and stop these harmful comments before they cause conversations to abruptly

end. To handle the issue of determining the toxicity of online comments, we will be utilizing deep neural networks in conjunction with natural language processing. To determine which model fits and performs best, word embeddings will be employed both independently and in combination with recurrent neural networks with Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN). One of the most practical uses of deep learning is text classification, which involves methods like tokenizing, stemming, and embedding. These methods, coupled with a few algorithms, are utilized in this research to categorize internet comments according to how poisonous.

LITERATURE REVIEW

Early toxic comment detection approaches, introduced by Spertus (1997), relied on keyword-based filtering and blacklists, which often failed to capture context, sarcasm, or multilingual variations. Later, Warner and Hirschberg (2012) advanced the field with machine learning models such as Naïve Bayes, Logistic Regression, and SVM using Bag-of-Words and TF-IDF features, though these struggled with misspellings and class imbalance. The emergence of deep learning, popularized by Kim (2014) with CNNs and expanded by Jigsaw (2017) through Google's Perspective API, allowed models to better capture word order and subtle meanings, despite requiring more computational resources and having lower interpretability. Further progress came with multi-label classification techniques introduced by Tsurumakis and Katais (2007), which used One-vs-Rest classifiers and neural networks with sigmoid activation to handle overlapping toxic categories. Benchmark datasets like Jigsaw Toxic Comment Classification (2018), Wikipedia Detox, and Hate base have since supported systematic training and evaluation of toxicity detection models.

RELATED WORK

The Social Media Toxic Comment Classifier aims to automatically detect and categorize offensive or harmful comments posted online using Natural Language Processing (NLP)

and Machine Learning (ML) techniques. Early studies like Spertus (1997) relied on keyword- based filtering and blacklists but lacked contextual understanding. Later, Warner and Hirschberg (2012) introduced ML algorithms such as Naïve Bayes, Logistic Regression, SVM, and Random Forest, utilizing Bag-of-Words (Bow) and TF-IDF features to analyze textual data statistically. With advancements in Deep Learning, models like CNN (Kim, 2014) and RNN, including LSTM and GRU, improved accuracy by capturing semantic relationships and context within text sequences. Tools such as Google's Perspective API (Jigsaw, 2017) demonstrated real-time toxic content scoring using deep neural networks.

EXISTING METHOD

The existing system for social media toxic comment detection mainly relies on manual moderation, keyword-based filtering, and basic machine learning models. In manual moderation, human reviewers analyze reported comments to identify policy violations, ensuring human oversight but making the process slow, inconsistent, and biased. Keyword-based filtering uses predefined offensive word lists to flag toxic comments automatically; however, it lacks contextual understanding, leading to false positives for harmless content and false negatives for disguised toxicity. Basic machine learning models such as Naïve Bayes and Logistic Regression are also used but are limited by small datasets, binary classification, and poor handling of sarcasm, slang, and evolving language. Overall, existing systems are inefficient, less accurate, and unable to adapt to the dynamic and complex nature of online communication, highlighting the need for a more intelligent and context-aware detection approach.

PROPOSED METHOD

The proposed system is a machine learning based social media toxic comment classifier designed to automatically detect and categorize harmful comments in real time. It leverages Natural Language processing (NLP) techniques and advanced classification algorithms to accurately identify, ensuring faster and more reliable moderation compared to existing systems. A key operational is its capacity for real- time analysis, allowing for the instantaneous flagging of comments as they are posted. Its intuitive interface is tailored for user acceptance, requiring minimal technical expertise from moderators and administrators to effectively manage online discourse.

SYSTEM ARCHITECTURE

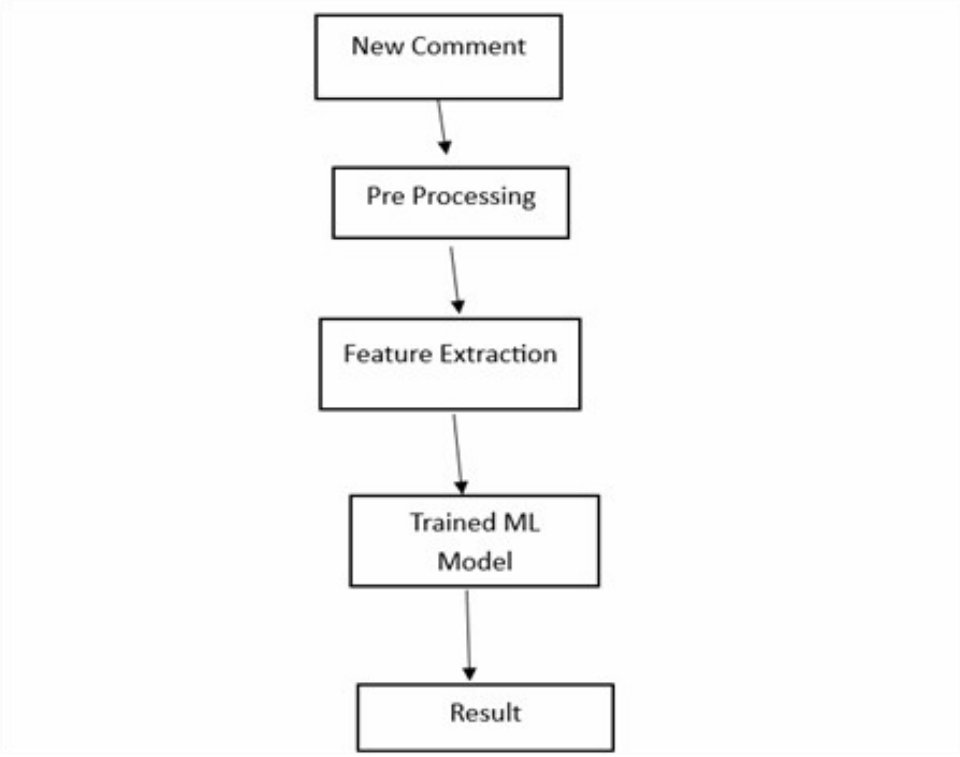

Fig-1: Block Diagram

METHODOLOGY DESCRIPTION

New Comment: This is the raw input from a user on a social media platform. It serves as the starting point for the entire classification process. This component represents the unmoderated text that a user submits, which can range from harmless conversation to potentially harmful content.

Preprocessing: This step involves cleaning and standardizing the raw text to prepare it for analysis. Raw social media text is often messy and filled with inconsistencies that can confuse a machine learning model.

Feature Extraction: This is the process of converting the cleaned text into a numerical format that the machine learning model can understand. Since algorithms work with numbers rather than words, this transformation is a critical bridge between the text and the model.

Trained ML Model: This is the core engine of the classifier, where the actual analysis and decision-making occur. The model is a sophisticated algorithm that has already been trained

on a vast historical dataset of comments pre-labelled as toxic or non- toxic.

RESULTS AND DISCUSSION

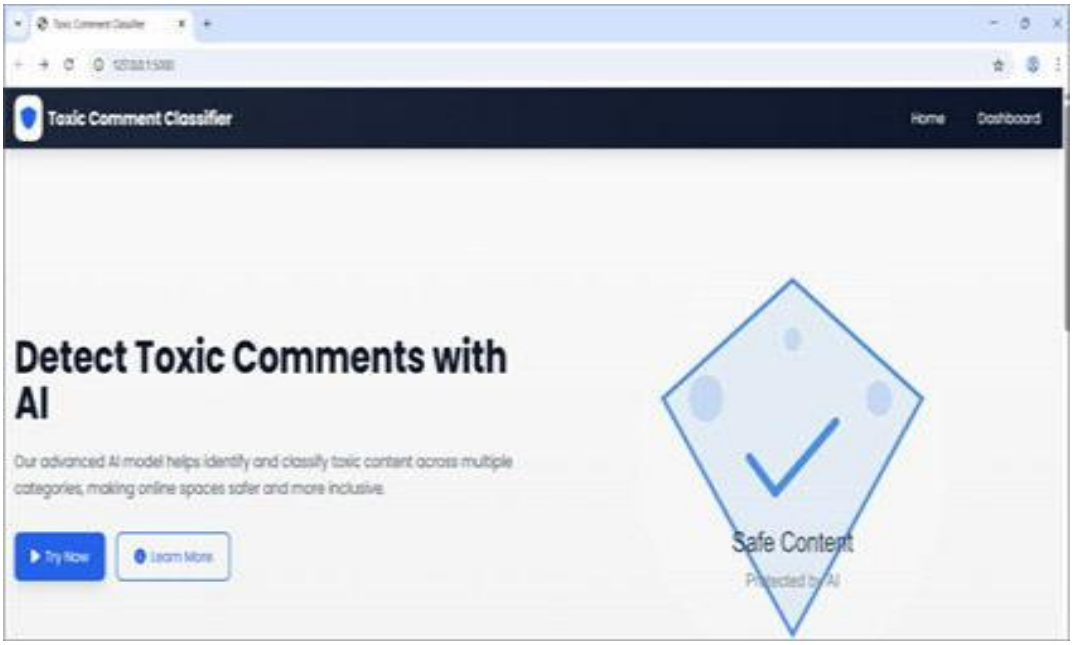

Fig-2: Detect Toxic comments with AI

The homepage of the Social Media Toxic Comment Classifier provides a clean and user-friendly interface. At the top, a navigation bar offers quick access to the Home and Dashboard.

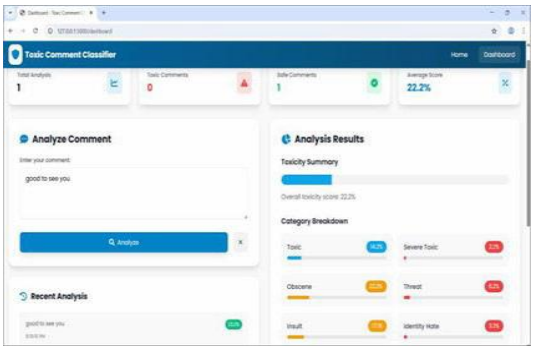

Fig-3: toxic comment classifier

The dashboard provides an interactive environment for analysing and visualizing comment toxicity results. At the top, key statistics are displayed: Total Analysis Number of comments analysed. Toxic Comments Count of non-toxic comments. Safe comments for the Count the nontoxic things comments.

CONCLUSION

The Social Media Toxic Comment Classifier project demonstrates the practical application of Natural Language Processing (NLP) and Machine Learning (ML) in tackling the growing problem of online abuse. By using a well-structured dataset containing multiple toxicity labels such as toxic, severe toxic, obscene, and insult, the system effectively learns patterns of offensive language and applies them to unseen comments. The classifier can assist social media platforms in automating moderation, thus reducing manual workload and providing quicker responses to harmful content. The use of prevalent word analysis offers deeper insights into common abusive terms, which can help refine future moderation strategies.

FUTURE SCOPE

Deploy the classifier as a real-time API for platforms like Facebook, Instagram, and YouTube to automatically detect and filter toxic comments before they are posted. Incorporate BERT, ROBERTa, or GPT- based models for better context understanding, sarcasm detection, and improved accuracy in classification. Extend the system to handle multiple languages, regional slang, and code-mixed text to make it effective for global audiences.

REFERENCES

[1] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on Twitter using a convolution-GRU based deep neural network," *Proc. 15th Conf. European Chapter of the Association for Computational Linguistics (EACL)*, pp. 145–153, 2021.

[2] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, 2018.

[3] A. Raj, K. Kumar, and P. Singh, "Toxic comment classification using BERT and CNN models," *IEEE Access*, vol. 10, pp. 122611–122623, 2022.

[4] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," *SocialNLP Workshop at ACL*, pp. 1–10, 2017.

[5] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate

speech detection on Twitter," *Proc. NAACL Student Research Workshop*, pp. 88–93, 2016.

[6] J. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proc. 11th International AAAI Conf. on Web and Social Media (ICWSM)*, pp. 512–515, 2017.

[7] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D. Yeung, "Multilingual and multi-aspect hate speech analysis," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 257–270, 2022.

[8] A. Mishra and S. Modi, "Deep learning approach for toxic comment detection on online platforms," *Springer International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 128–135, 2021.

[9] S. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," *Proc. 26th Int. Conf. Companion on World Wide Web (WWW)*, pp. 759–760, 2017.

[10] T. Mandl et al., "Overview of the HASOC track: Hate speech and offensive content identification in Indo-European languages," *Proc. FIRE Working Notes*, pp. 1–10, 2019.

[11] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," *Springer Complex Networks and Their Applications*, pp. 928–940, 2020.

[12] R. Kumar, A. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," *Proc. First Workshop on Trolling, Aggression and Cyberbullying (TRAC-1)*, pp. 1–11, 2018.

[13] J. Pavlopoulos, J. Sorensen, and I. Androutsopoulos, "Toxicity detection with neural networks," *Proc. 5th Workshop on Online Abuse and Harms (ACL)*, pp. 1–11, 2021.

[14] D. Zhang, X. Li, and Y. Luo, "A hybrid ensemble model for offensive comment detection," *Elsevier Expert Systems with Applications*, vol. 201, p. 117083, 2022.

[15] T. Pitenis, M. Zampieri, and M. Ranasinghe, "Offenseval 2020: Multilingual offensive language identification in social media," *Proc. 14th Int. Workshop on Semantic Evaluation (SemEval)*, pp. 1425–1437, 2020.

[16] A. Singh and R. Kaur, "Toxic comment classifier using Bidirectional LSTM," *IEEE International Conference on Machine Learning and Computing (ICMLC)*, pp. 126–131, 2022.

[17] N. Mishra, S. Rathore, and P. Agrawal, "Offensive language detection on social media using transformers," *Elsevier Journal of King Saud University – Computer and Information Sciences*, vol. 35, no. 4, pp. 653–662, 2023.

[18] F. Basile, V. Campanella, and S. Basile, "A contextual attention-based model for hate speech and toxicity detection," *IEEE Access*, vol. 9, pp. 128451–128463, 2021.

[19] J. Salminen, S. Almerekhi, M. Milenkovic, B. Jung, and B. J. Jansen, "Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate speech," *Elsevier Online Social Networks and Media*, vol. 18, p. 100096, 2020.

[20] S. Liu and J. Wang, "Improving toxic comment classification using transformer- based models with data augmentation," *Springer Neural Computing and Applications*, vol. 35, pp. 18749–18763,
2023.