



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

TOXIC COMMENT CLASSIFICATION

¹ Harshad M. Kubade

Guide, Department of Information Technology,
Priyadarshini College of Engineering, Nagpur, India
harshad.kubade@pce.edu.in

² Rewati A. Kawale

UG Student, Department of Information Technology,
Priyadarshini College of Engineering, Nagpur, India
rewatikawale65@gmail.com

³ Ishika S. Kahar

UG Student, Department of Information Technology,
Priyadarshini College of Engineering, Nagpur, India
ikahar99@gmail.com

⁴ Jatin P. Bais

UG Student, Department of Information Technology,
Priyadarshini College of Engineering, Nagpur, India
jatinbais3@gmail.com

⁵ Adarsh Anandrao Nimkar

UG Student, Department of Information Technology,
Priyadarshini College of Engineering, Nagpur, India
adarshnimkar04@gmail.com

⁵ Manish Vasanta Dhoble

UG Student, Department of Information Technology,
Priyadarshini College of Engineering, Nagpur, India
vasantadhoble143@gmail.com

Abstract – The advent of social media and online platforms has brought about an unprecedented surge in user-generated content. However, with the freedom of expression also comes the challenge of combating toxic behavior, such as hate speech, harassment, and offensive comments, which can significantly impact online discourse and community well-being. In response to this challenge, this study proposes a machine-learning approach for classifying toxic comments. The primary objective of this research is to develop an effective and efficient model for automatically identifying toxic comments within large volumes of user-generated content. Leveraging a diverse dataset of comments labeled with toxicity levels, various machine learning algorithms, including but not limited to logistic regression, support vector machines, and neural networks, are explored and compared for their performance in toxic comment classification.

Keywords: Toxic Comment Classification, Machine Learning, Natural Language Processing, Text Classification, Hate Speech Detection.

I. INTRODUCTION

In the digital age, online platforms serve as powerful mediums for communication, collaboration, and expression. However, alongside the vast opportunities for interaction and discourse, the proliferation of toxic behavior in online communities has emerged as a pressing concern. Toxic comments, encompassing hate speech, harassment, insults, and other forms of harmful content, not only undermine the quality of online interactions but also pose significant risks to individual well-being and community cohesion.

The classification of toxic comments has therefore garnered substantial attention from researchers, platform moderators, and policymakers alike. By automating the identification and moderation of toxic content, machine learning techniques offer a promising avenue for mitigating the adverse effects of toxic behavior and fostering healthier online environments.

This introduction sets the stage for understanding the significance of toxic comment classification and outlines the objectives, challenges, and methodologies underlying this research endeavor.

Objectives:

The primary objective of toxic comment classification is to develop computational models capable of automatically discerning between toxic and non-toxic comments within large volumes of user-generated content. By accurately identifying toxic behavior, these models enable platform moderators to efficiently moderate discussions, protect users from harassment, and maintain a positive online environment.

Furthermore, toxic comment classification serves broader societal goals, including the promotion of free expression, the prevention of cyberbullying, and the preservation of online civility. Through systematic analysis and categorization of toxic comments, researchers can gain insights into the prevalence, dynamics, and underlying drivers of toxic behavior, informing interventions and policy decisions aimed at fostering safer and more inclusive online spaces.

Challenges:

Toxic comment classification presents several challenges stemming from the nuanced nature of language, the diversity of toxic behaviors, and the evolving landscape of online communication. Key challenges include:

Linguistic Complexity: Toxic comments exhibit diverse linguistic patterns, including sarcasm, irony, and cultural nuances, making it challenging to develop robust classification models capable of capturing subtle cues indicative of toxicity.

Data Imbalance: Toxic comments are often rare compared to non-toxic comments, resulting in imbalanced datasets that can bias model performance and hinder generalization to real-world scenarios.

Context Sensitivity: The interpretation of toxicity varies across different contexts, communities, and cultural norms, necessitating adaptable classification strategies that account for contextual nuances and user perceptions.

Model Interpretability: As automated moderation decisions can have significant implications for user experience and freedom of expression, ensuring the transparency and interpretability of classification models is essential to foster trust and accountability.

II. LITERATURE SURVEY

Rahul, H. Kajla, et al., [1], toxic comments, characterized by their disrespectful, abusive, or unreasonable nature, often drive users away from online discussions. The pervasive threat of online bullying and harassment has a chilling effect on the expression of dissenting viewpoints, hampering the free exchange of ideas. Consequently, websites face

challenges in fostering productive discussions, prompting many communities to impose restrictions or shut down user comments entirely. This paper aims to conduct a systematic analysis of online harassment, endeavoring to accurately classify its various manifestations to better understand and address the toxicity prevalent in online interactions.

A. Garlapati et al., [2], this study introduces an innovative application of Natural Language Processing (NLP) techniques to classify the nature of toxicity in online comments. By employing advanced NLP methodologies, our analysis seeks to identify and categorize different types of toxic comments, including but not limited to those categorized as obscene, identity-based hate, threatening, insulting, and severely toxic. Leveraging comments sourced from online platforms, our algorithm is trained to discern between toxic and non-toxic content, ultimately aiming to predict the specific toxicity class associated with each comment.

Ashish et al., [3], the primary goal of this study is to enhance the effectiveness and precision of toxic comment classification by tackling existing challenges. Specifically, our research endeavors to refine the model's capacity to identify and categorize nuanced expressions of toxic language. To achieve this, we propose incorporating supplementary contextual cues and employing advanced techniques like adversarial training and data augmentation to enrich the diversity of training data. Additionally, our study aims to assess the model's performance across various real-world datasets to validate its applicability in practical scenarios.

F. Museng et al., [4], results reveal that Long Short-Term Memory (LSTM) emerges as the most frequently referenced deep learning model across the examined research papers, appearing in 8 out of 26 instances. Moreover, LSTM consistently demonstrates impressive accuracy rates, surpassing 79% in multiple studies, particularly when trained on datasets of approximately 9000 samples. Notably, the performance of LSTM models may vary depending on the pre-processing techniques employed.

V. Swetha et al., [5], the study aims to meticulously assess the prevalence of online harassment and categorize the content into distinct labels to scrutinize its toxicity comprehensively. To achieve this, six machine learning algorithms are employed and applied to our dataset to address the challenge of text classification. The objective is to identify the most effective machine learning algorithm based on our evaluation metrics for accurately classifying harmful remarks.

M. Husnain et al., [6], the study employs an innovative preprocessing approach that converts the multi-label classification task into a multi-class classification task. This preprocessing strategy has demonstrated notable enhancements in accuracy when applied to basic classification models, thereby advocating for its adoption in more advanced models as well.

III. METHODOLOGY

Toxic comments pose a significant challenge to online communities, undermining healthy discourse and fostering an environment of hostility. In response to this pressing issue, we propose a machine learning-based system for classifying toxic comments. Leveraging advanced algorithms and techniques, our system aims to accurately identify and categorize toxic comments, thereby facilitating proactive moderation and promoting a more respectful online environment.

Introduction: Toxic comments pose a significant challenge to online communities, undermining healthy discourse and fostering an environment of hostility. In response to this pressing issue, we propose a machine learning-based system for classifying toxic comments. Leveraging advanced algorithms and techniques, our system aims to accurately identify and categorize toxic comments, thereby facilitating proactive moderation and promoting a more respectful online environment.

System Components:

Data Collection and Preprocessing:

- Acquire a diverse dataset of user comments from various online platforms.
- Preprocess the data to remove noise, normalize text, and handle imbalances in the dataset.

Feature Extraction:

- Extract relevant features from comment texts, such as word embeddings, n-grams, and syntactic structures.
- Utilize techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to represent the textual features effectively.

Model Development:

- Implement a variety of machine learning algorithms for toxic comment classification, including logistic regression, support vector

machines, random forests, and neural networks.

- Fine-tune hyperparameters and explore ensemble methods to improve model performance.

Training and Evaluation:

- Train the classification models on the preprocessed dataset, using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score.
- Employ cross-validation techniques to validate the robustness of the models and mitigate overfitting.

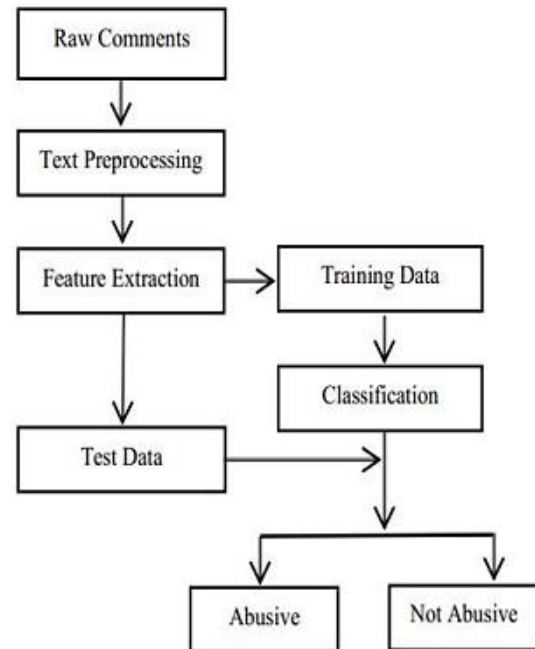


Figure 3.1 Block Diagram

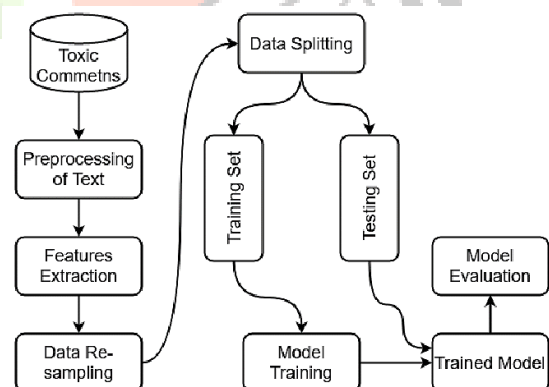
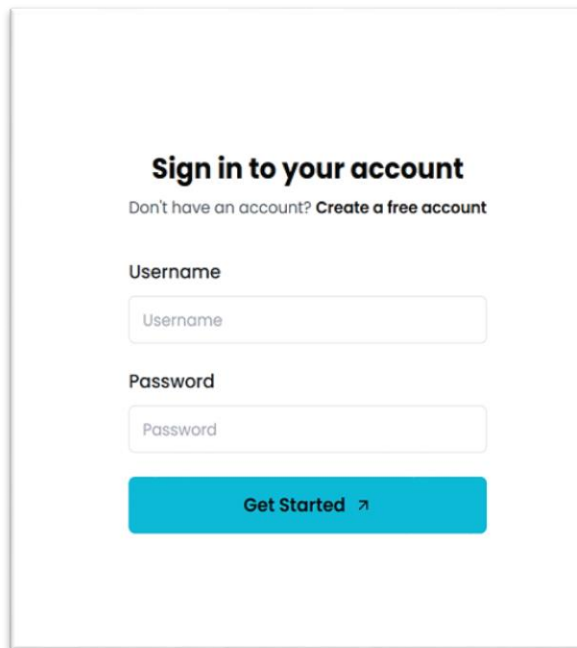


Figure 3.2 Flow Diagram



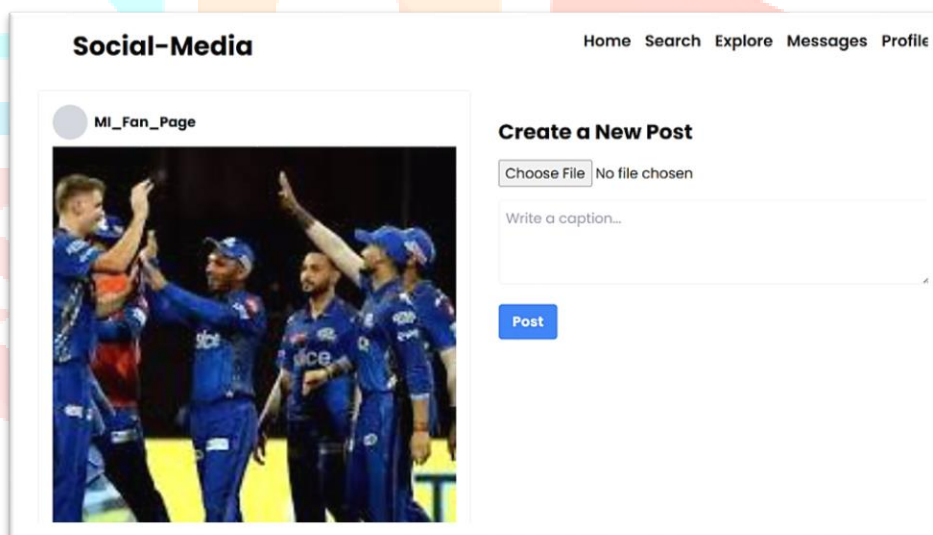
Sign in to your account
Don't have an account? [Create a free account](#)

Username

Password


[Get Started ↗](#)

Figure 4.1 Registration/Login Form



Social-Media Home Search Explore Messages Profile

MI_Fan_Page



Create a New Post
 No file chosen
Write a caption...

Figure 4.2 Social Media Dashboard Form

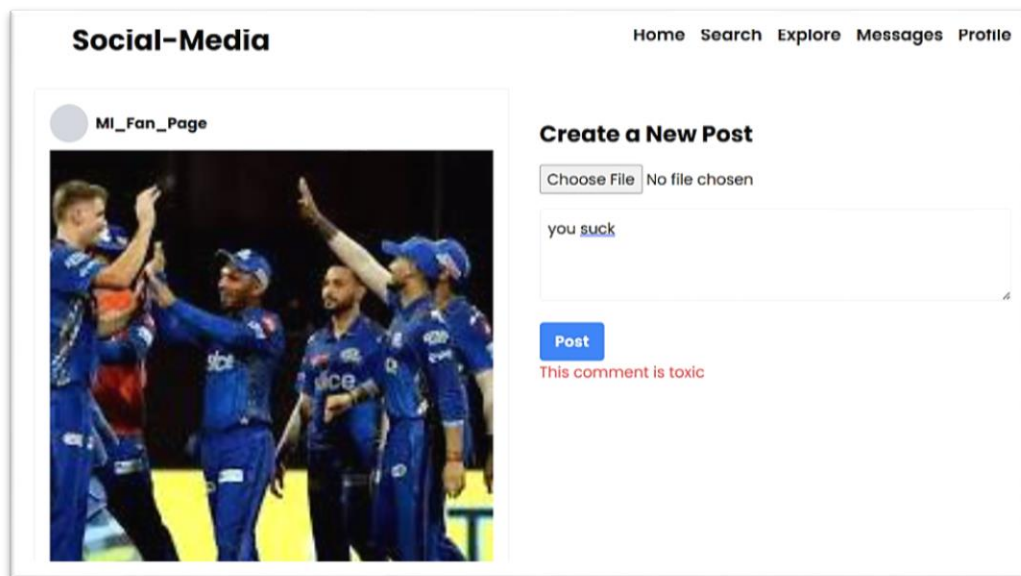


Figure 4.3 Comment Classification

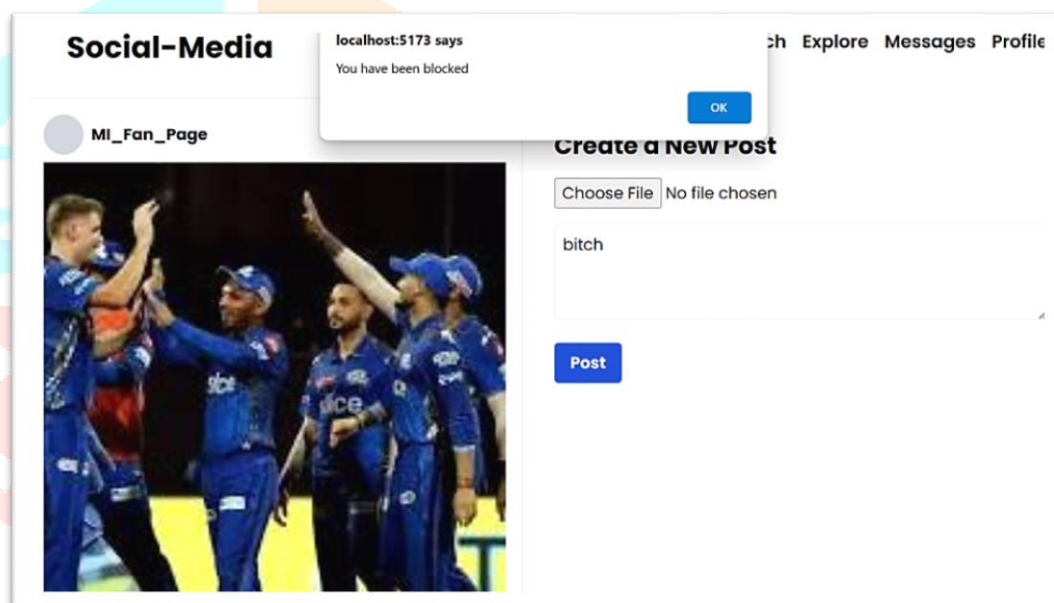


Figure 4.4 Toxic Comment Classified

In the above shown results of the Toxic Comment Classification (TCC) project we developed a system that classifies user comment on a particular image like on other social media platforms we use. In this project, we have given open access for everyone to upload images or pass any comment on it so that the given comment can be classified based on the behavior of that word. If the comment is normal then no action is taken and if something is wrong or a bad comment then the user gets blocked from the site and cannot use it anymore. The idea behind this was to make a project that can not only give social media accessing feeling but also secure us from bad, offensive, or hurting comments we face during social media access.

V. CONCLUSION

In conclusion, the application of machine learning techniques for toxic comment classification represents a significant step towards fostering safer and more inclusive online communities. Through the development and deployment of advanced classification models, this approach enables platforms to effectively identify and mitigate toxic behavior, thereby enhancing user experience and promoting constructive engagement. Overall, toxic comment classification using machine learning holds immense promise in fostering healthier online interactions, safeguarding user well-being, and upholding the principles of free expression and diversity of viewpoints in the digital age. Through continued collaboration between researchers, platform moderators, and policymakers, we can work towards creating a more respectful and inclusive online environment for all users.

VI. RREFERENCES

- [1] Rahul, H. Kajla, J. Hooda and G. Saini, "Classification of Online Toxic Comments Using Machine Learning Algorithms," *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020, pp. 1119-1123, doi: 10.1109/ICICCS48265.2020.9120939.
- [2] Garlapati, N. Malisetty and G. Narayanan, "Classification of Toxicity in Comments using NLP and LSTM," *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2022, pp. 16-21, doi: 10.1109/ICACCS54159.2022.9785067.
- [3] Ashish, A. Rani, and H. Shyan, "A Comparative Study and Analysis on Toxic Comment Classification," *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2023, pp. 783-787, doi: 10.1109/ICSCSS57650.2023.10169771.
- [4] F. Museng, A. Jessica, N. Wijaya, A. Anderies and I. A. Iswanto, "Systematic Literature Review: Toxic Comment Classification," *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, Yogyakarta, Indonesia, 2022, pp. 1-7, doi: 10.1109/ICITDA55840.2022.9971338.
- [5] V. Swetha, R. Anuhya, E. S. Sowmya and A. Geethanjali, "Building a Toxic Comments Classification Model," *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2021, pp. 1519-1523, doi: 10.1109/ICECA52323.2021.9675911.
- [6] M. Husnain, A. Khalid and N. Shafi, "A Novel Preprocessing Technique for Toxic Comment Classification," *2021 International Conference on Artificial Intelligence (ICAI)*, Islamabad, Pakistan, 2021, pp. 22-27, doi: 10.1109/ICAI52203.2021.9445252.
- [7] K. Dubey, R. Nair, M. U. Khan and P. S. Shaikh, "Toxic Comment Detection using LSTM," *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, Bengaluru, India, 2020, pp. 1-8, doi: 10.1109/ICAECC50550.2020.9339521.
- [8] Gurshobit Singh Brar and Ankit Sharma, Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques, Nov 2018.
- [9] E. Wulczyn, N. Thain and L. Dixon, "Ex Machina: Personal attacks are seen at scale", *Proceedings of the 26th International Conference on World Wide Web*, pp. 1391-1399, 2017.
- [10] L. Song, R. Y. Lau, and C. Yin, "Discriminative Topic Mining for Social Spam Detection", *PACIS*, pp. 378, 2014

