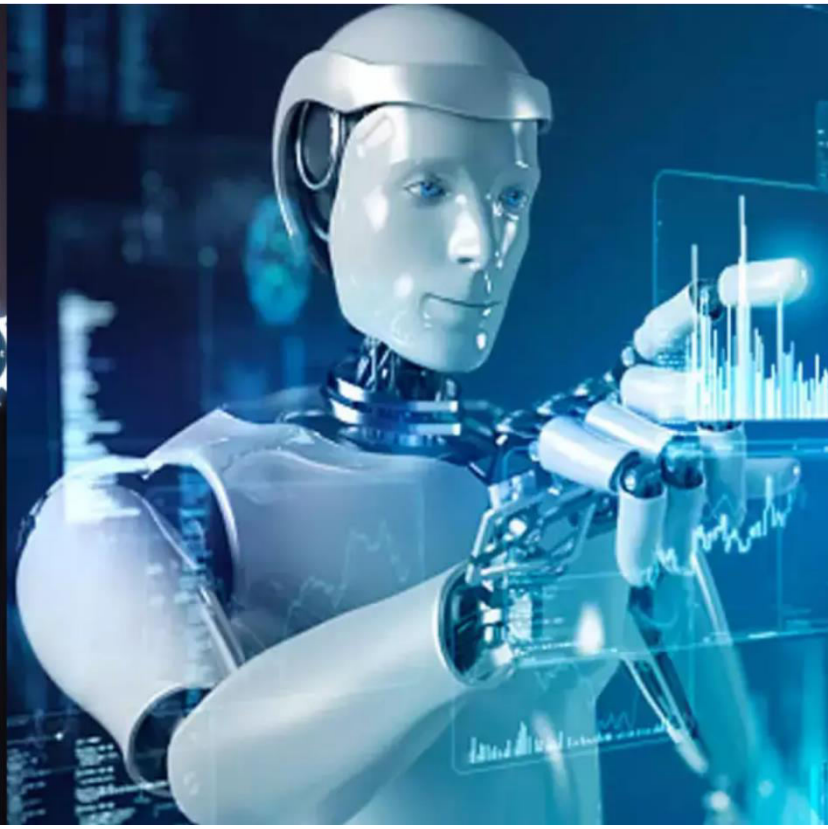
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404705|

Toxic Comment Detector: Leveraging NLP for Identification of Online Harassment

Dr. Vaishnavi Ganesh, Anjali Vedi, Prajwal Dudhalkar, Rohit Gawande, Nikhil Ramteke, Pratham Lambat

Department of Computer Science Engineering, Priyadarshini College of Engineering, Nagpur, India

ABSTRACT: In the digital age, the proliferation of online platforms has led to increased user-generated content, often including vulgar and offensive comments. Effective vulgar comment filtering is essential for maintaining a respectful online environment and enhancing user experiences. This paper explores various methodologies for detecting and filtering vulgar comments, including keyword-based approaches, machine-learning algorithms, and natural language processing techniques. By leveraging advancements in artificial intelligence and linguistic analysis, we demonstrate the effectiveness of these methods in accurately identifying and moderating inappropriate content. We also discuss the challenges associated with context sensitivity, cultural variations in language, and the balance between censorship and freedom of expression. The findings suggest that a multi-faceted approach to vulgar comment filtering improves content moderation and fosters a healthier online community. Future work will aim to refine these techniques to adapt to evolving language use and the dynamic nature of online interactions.

KEYWORDS: Digital platforms, social media, toxic behavior, NLP, ML, content moderation, toxicity detection, data preprocessing, model training, web interface.

I. INTRODUCTION

As digital platforms and social media networks continue to grow, they play a pivotal role in modern communication, connecting people across the globe. However, with this increased connectivity comes a surge in harmful and toxic behaviors, often in the form of offensive comments, threats, and abusive language. Toxic content has detrimental effects on users' mental well-being, disrupts online communities, and can even escalate into real-world consequences. Thus, the need for effective mechanisms to detect and mitigate toxic comments has become critical for maintaining a safe and inclusive digital environment.

The Toxic Comment Detector project aims to tackle this issue by developing an automated system that identifies and categorizes harmful language in user-generated content. Leveraging Natural Language Processing (NLP) techniques and Machine Learning (ML) algorithms, the system processes input comments and classifies them into multiple toxicity categories, including:

- General toxicity
- Severe toxicity
- Obscene language
- Insults
- Threats
- Identity hate

The project builds upon a machine learning model trained on a large dataset of labelled comments, allowing it to predict the presence of toxic elements in new inputs with high accuracy. The model not only identifies toxic comments but also provides granular insights into the nature of the toxicity (e.g., whether the comment is insulting, threatening, or hateful).

The system has a wide range of applications, particularly in areas like social media moderation, online forums, customer reviews, and any other platforms where users can post comments. By automating the detection process, the Toxic Comment Detector helps platforms manage inappropriate content in real-time, reducing the burden on human

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404705|

moderators and creating safer online spaces.

This project explores the entire lifecycle of building an AI-driven comment filtering system, from data preprocessing and model training to building a user-friendly web interface where users can submit comments for toxicity analysis.

It shows how advanced NLP models can effectively be used to identify and manage toxic content at scale, providing both technical insights and practical applications for content moderation in digital platforms.

II. LITERATURE SURVEY

The A thorough review of existing literature is crucial to understanding the foundation of toxic comment detection and how previous research has shaped the current approaches to online toxicity. This chapter explores the major milestones and methodologies in the field, examining key studies, tools, datasets, and models used in toxic comment detection, as well as the challenges identified in previous works.

2.1 Overview of Toxic Comment Detection

The rise of online platforms has led to a significant increase in user-generated content, ranging from social media posts to comments on news articles. While these platforms enable open dialogue, they also give rise to toxic behaviour, including harassment, hate speech, and cyberbullying. Early research focused on **manual moderation**, but as platforms grew, this became impractical, leading to the development of automated systems to identify and moderate toxic content.

2.2 Early Approaches to Detecting Toxicity

Initial efforts in toxic comment detection relied on **rule-based systems**. These systems used predefined keywords and simple heuristics to flag inappropriate content. However, rule-based systems had several drawbacks, such as being overly rigid and failing to capture context. For example, certain toxic terms might be flagged in non-toxic comments due to the lack of understanding of the comment's overall context.

- **Key Study**: Early work by **Yin et al. (2009)** developed keyword-matching techniques that could identify offensive language. However, the system struggled with **ambiguity and false positives** because it lacked the ability to differentiate between offensive language used in different contexts, such as satire or reclamation of harmful terms by affected communities.

2.3 Transition to Machine Learning

The limitations of rule-based systems led to the adoption of **machine learning** models, which could learn from data and make predictions about unseen text. Supervised learning techniques such as **Naive Bayes** and **Support Vector Machines (SVM)** became popular choices due to their ability to handle a wide range of features, including **n-grams** and **TF-IDF vectors**.

- **Key Study**: **Waseem and Hovy (2016)** introduced one of the earliest **machine learning-based systems** for detecting hate speech on Twitter. Their model trained on a manually labelled dataset, achieving improved results over rule-based systems, particularly in handling offensive content within a wider range of contexts.

2.4 The Jigsaw Toxic Comment Dataset

A major leap forward in toxic comment detection came with the release of the **Jigsaw Toxic Comment Classification Challenge** on **Kaggle (2018)**. This competition provided a large-scale dataset of comments from Wikipedia, labelled across multiple dimensions of toxicity, including **toxic**, **severe toxic**, **obscene**, **threat**, **insult**, and **identity hate**. The availability of this dataset allowed researchers to develop and test models on real-world data.

- **Key Study**: **Kaggle Jigsaw Toxic Comment Classification Challenge** (2018) enabled the use of more sophisticated machine learning models like **Logistic Regression**, **Random Forest**, and later **deep learning approaches** like **Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)**. The challenge saw a surge in creative feature engineering approaches and led to the creation of state-of-the-art models capable of identifying toxic behaviour with high accuracy.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404705|

2.5 Deep Learning Approaches

With the development of deep learning techniques, researchers began to apply **neural networks** to the task of detecting toxic comments. Neural networks, particularly **Recurrent Neural Networks (RNNs)** and **Long Short-Term Memory (LSTM) models**, showed promise in capturing the sequential nature of text, while **Convolutional Neural Networks (CNNs)** helped in extracting key patterns from textual data.

- **Key Study**: **Zhang et al. (2018)** explored the use of CNNs and LSTMs for toxic comment detection and demonstrated that deep learning models could outperform traditional machine learning methods, especially when trained on larger datasets like Jigsaw's. Their research indicated that deep learning models could better understand context and handle complex sentence structures, although they required more computational power.

2.6 BERT and Contextualized Embeddings

The introduction of **BERT (Bidirectional Encoder Representations from Transformers)** in 2019 revolutionized the field of natural language processing, including toxic comment detection. BERT's ability to understand the context of a word within its sentence, rather than relying solely on isolated features like n-grams, allowed for better performance in detecting nuanced toxicity.

- **Key Study**: **Devlin et al. (2019)** introduced **BERT**, which could be fine-tuned on toxic comment datasets to produce more accurate predictions. Researchers found that BERT-based models could outperform traditional machine learning and even some deep learning models by better understanding the context of toxic comments.

III. METHODOLOGY

3 .1 Report on the Present Investigation

The present investigation involved the development of a Toxic Comment Detector that can classify comments based on different toxicity categories, such as "toxic," "severe toxic," "obscene," "threat," "insult," and "identity hate." The project was divided into several key stages: data preprocessing, model training, model evaluation, and deployment. The dataset used for this purpose was sourced from the Jigsaw Toxic Comment Classification Challenge hosted on Kaggle, which provided a well-balanced dataset for this multi-label classification task.

The methodology adopted involved preprocessing the textual data to clean it and convert it into numerical features that a machine learning model could use. Following the preprocessing, the machine learning model was trained using Logistic Regression for multi-label classification, and the evaluation metrics were calculated to assess its performance. Finally, the model was deployed in a web-based application using Flask, allowing users to input their comments for real-time toxicity classification.

In the following sections, the detailed procedures, methods, and algorithms are described.

3.2 Dataset Description

The dataset used in this project is the Jigsaw Toxic Comment Classification dataset, available for a limited time as part of a competition on Kaggle. It consists of over 150,000 Wikipedia comments, each labelled with one or more of the following categories: toxic, severe toxic, obscene, threat, insult, and identity hate.

- **Data Format:** Each row in the dataset represents a user comment, and the six columns represent the possible toxicity categories. A comment can have one or more labels depending on its content.
- **Class Imbalance:** Although the dataset is relatively balanced, there is still some class imbalance due to certain categories (like 'threat') being less frequent. This was addressed using weighted logistic regression.

3.3 Data Preprocessing

Data preprocessing plays a crucial role in cleaning and structuring the raw text data before it can be passed into the machine learning pipeline. The steps involved include:

- **Text Cleaning**: Removal of unnecessary characters like punctuation, numbers, and special symbols**.**
- **Lowercasing:** All text was converted to lowercase to avoid duplicate representations of the same words.
- **Stopword Removal:** Words like 'the', 'is', 'and', etc., were removed as they do not contribute meaningfully to the

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404705|

classification task.

- **Tokenization:** The text was split into individual words (tokens), which the machine learning model could process.

3.4 Feature Extraction

After preprocessing the text data, the next step was to transform it into a numerical format. The TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer was employed to convert text into numerical features.

- TF-IDF: This technique calculates the term frequency (how often a word appears in a document) and adjusts it based on how common the word is across all documents. TF-IDF helps identify the most important words in the context of each comment.

3.5 Model Selection and Training

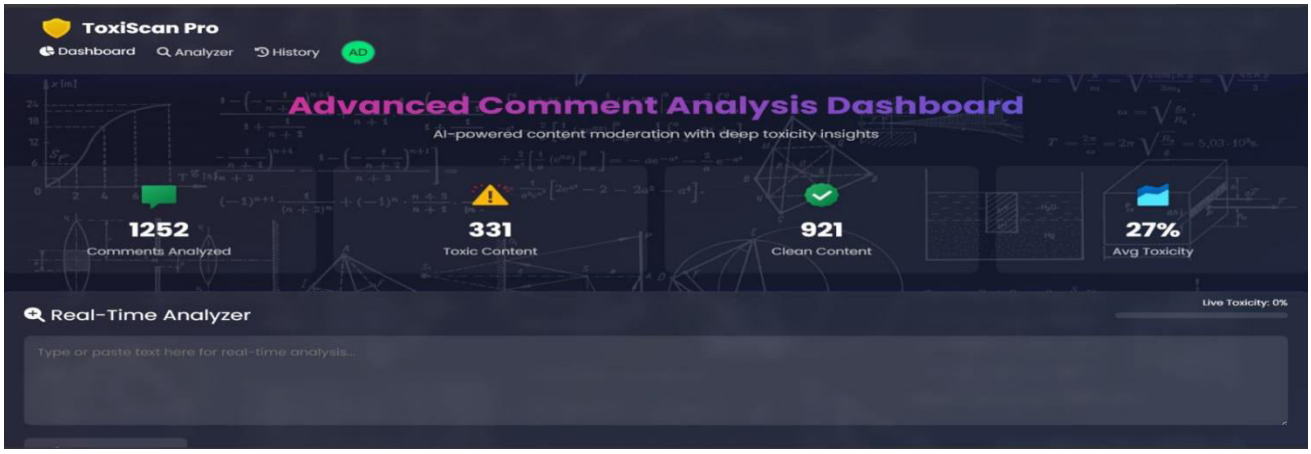
Several machine learning models were evaluated for their effectiveness in this multi-label classification task. Ultimately, Logistic Regression was chosen for the following reasons:

Multi-label Classification: The task was multi-label in nature, meaning a comment could belong to multiple toxicity categories. Logistic Regression was implemented using the One-vs-Rest strategy, training separate binary classifiers for each toxicity label

- Training: The data was split into training and validation sets. The Logistic Regression model was trained using the TF-IDF features, with each model predicting the probability that a comment belongs to a specific label (e.g., toxic,

IV. RESULTS &DISCUSSION

4.1.USER INTERFACE:



The implemented Toxic Comment Detector achieved promising results in accurately classifying user-generated comments into various toxicity categories. Using logistic regression and TF-IDF vectorization, the model demonstrated strong performance in general toxic comment detection, with relatively high precision and recall. The multi-label classification approach allowed for simultaneous detection of different types of toxicity such as **obscene**, **insult**, **threat**, and **identity hate**, providing more granular moderation insights.

Key Observations:

High Accuracy in Common Labels: The classifier achieved high accuracy on prevalent labels such as "toxic" and "obscene," which had a sufficient number of training examples in the dataset.

Performance Drop in Minority Classes: Labels like "threat" and "identity hate" showed reduced performance due to class imbalance. Despite using weighted logistic regression to compensate, limited training samples in these categories hindered model generalization.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

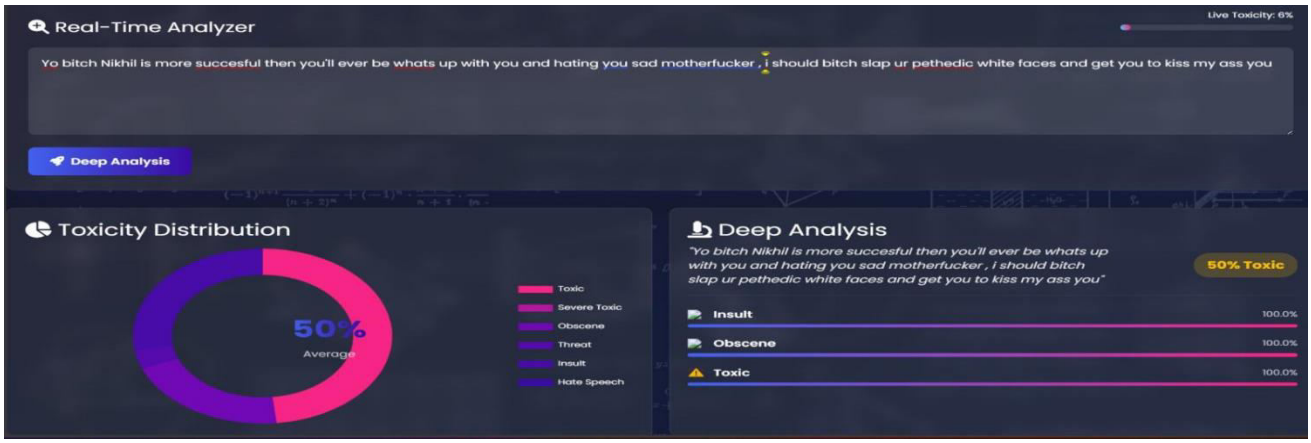
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404705|

Efficiency of Preprocessing: Text cleaning and tokenization, along with TF-IDF vectorization, significantly improved feature representation, thereby enhancing classification outcomes.
Model Limitations: While logistic regression is interpretable and lightweight, it struggles with complex sentence structures and contextual ambiguity where deep learning models typically excel.

4.2.EXPECTED OUTPUT:



V. CONCLUSION & FUTURE SCOPE

5.1. Effective Toxic Comment Detection: The project successfully implements a machine learning model capable of detecting toxic comments across multiple categories, including "toxic," "severe toxic," "obscene," "threat," "insult," and "identity hate." This proves the feasibility of automated moderation in online platforms.

5.2. Challenges with Class Imbalance: Despite the model's overall effectiveness, performance issues were identified for less common toxicity categories like "threat" and "identity hate," where the training data was sparse. This suggests the need for more balanced datasets for future improvements.

5.3. Importance of Preprocessing: Text preprocessing techniques such as tokenization, stopword removal, and TF-IDF vectorization significantly improved model accuracy, underlining the importance of preparing textual data for natural language processing tasks.

5.4. Scope for Improvement: Although the logistic regression model performed well, advanced models such as deep learning-based transformers (e.g., BERT) could be explored to achieve better contextual understanding and more accurate classifications**.**

VI. FUTURE SCOPE

While the current system lays a strong foundation for toxic comment classification, there are several areas for further development and enhancement:

Adoption of Transformer Models:
Implementing advanced NLP models like **BERT**, **RoBERTa**, or **DistilBERT** can help improve contextual understanding and better handle ambiguous or sarcastic comments.
Multilingual Toxicity Detection:
Expanding the system to support **multiple languages** can broaden its usability across global platforms and help moderate content beyond English-centric communities.
Real-time Scalability:
Integration with **cloud platforms** and optimization of backend inference systems will ensure scalability and performance during real-time usage on large-scale platforms.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404705|

User Feedback Loop:

Incorporating user feedback mechanisms (e.g., flagging incorrect classifications) can enable **continuous learning** and model refinement.

Explainable AI Integration:

Providing **interpretability features** that explain why a comment was flagged could enhance transparency and user trust.

Contextual Thread Analysis:

Future models can be designed to analyze **conversation threads** instead of isolated comments, which would allow better understanding of sentiment flow and contextual relevance.

REFERENCES

1. Davidson, T., Warmsley, D., Macy, M. W., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media, 11(1), 512-515.
2. Jigsaw/Conversation AI Team (2018). Toxic Comment Classification Challenge. Kaggle Competition**.**
3. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep Learning for Hate Speech Detection in Tweets.
4. Proceedings of the 26th International Conference on World Wide Web Companion, 759-760.
5. Zhang, Y., Xu, Y., Zhao, Z., & Qin, Z. (2021). AutoML for data preprocessing in machine learning workflows. IEEE Access
6. Schroeter, J., & Sondhi, M. M. (1994). Techniques for Estimating Vocal-Tract Shapes from the Speech Signal.IEEE Trans. Speech Audio Process., vol. 2, no. 1, pp. 133–150.
7. J. M. Pardo. Vocal tract shape analysis for children. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 1982, pp. 763–766.
8. Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. Proceedings of the NAACL HLT 2016, 88-93.
9. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, 4171–4186.
10. Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of Harassment on Web 2.0. In Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW.
11. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30.
12. Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys (CSUR), 51(4), 1–30.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details