International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Intelligent Toxic Comment Detection Using Machine Learning and Natural Language Processing Techniques

Dr.S.Suresh[1] , Namala Sireesha[2], Shaik Davud[3], Tirumani Bhanu Shankar Satyanarayana[4], Kada Rama Satya Pavan[5] , Kala Tirumala Venkata Sai Teja[6]

[1]Associate Professor, [2,3,4,5,6] B.tech Students Department of CSE, Pragati Engineering College, Surampalem , Andhra Pradesh, India

Abstract- — The rapid expansion of social media platforms and online communication systems has significantly increased the amount of user-generated content on the internet. While these platforms enable people to share ideas and communicate freely, they also expose users to harmful content such as hate speech, offensive language, cyberbullying, and abusive comments. Toxic comments not only affect healthy online discussions but also create negative psychological and social impacts on individuals. Therefore, developing automated systems capable of detecting and filtering toxic comments has become an important research problem in natural language processing and online content moderation. This study presents an intelligent framework for detecting toxic comments using machine learning and natural language processing techniques. The proposed system analyses textual data collected from online platforms and classifies comments into toxic and non-toxic categories. Various preprocessing techniques such as tokenization, stop-word removal, text normalization, and lemmatization are applied to clean and prepare the dataset for model training. Feature extraction methods including Term Frequency–Inverse Document Frequency (TF-IDF) and word embedding techniques are used to transform textual data into numerical representations suitable for machine learning models. Several machine learning and deep learning algorithms, including Naive Bayes, Support Vector Machines (SVM), Logistic Regression, and Convolutional Neural Networks (CNN), are implemented and compared to determine the most effective model for toxic comment classification. The models are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. Experimental results indicate that deep learning models, particularly CNN-based architectures, achieve higher classification accuracy and better performance in detecting complex toxic language patterns. The proposed system can assist online platforms in automatically identifying harmful content and maintaining safer digital communication environments. By integrating machine learning techniques with advanced natural language processing methods, the framework contributes to improving online content moderation and promoting respectful interactions in digital communities.

Index Terms: Toxic Comment Detection, Natural Language Processing, Machine Learning, Deep Learning, Text Classification, Cyberbullying Detection, Online Content Moderation.

I. INTRODUCTION

The rapid growth of the internet and social media platforms has significantly transformed the way people communicate and share information. Online platforms such as social networking sites, discussion forums, and messaging applications allow users to interact and exchange ideas instantly. However, alongside these benefits, the expansion of digital communication has also led to the widespread presence of harmful online content, including hate speech, abusive language, cyberbullying, threats, and offensive comments. Such toxic comments can negatively impact online communities and may cause emotional distress, psychological harm, and social conflicts among users. Consequently, maintaining a safe and respectful online environment has become a major challenge for social media platforms and online communities.

Manual moderation of online content is extremely difficult due to the massive volume of user-generated data produced every day. Social media platforms receive millions of comments and posts daily, making it nearly impossible for human moderators to efficiently review and filter all content. As a result, automated systems capable of detecting and filtering toxic comments have become essential for improving online safety and supporting effective content moderation.

Traditional approaches for detecting toxic comments mainly relied on rule-based systems and keyword filtering techniques. These systems typically use predefined dictionaries of offensive words and manually constructed rules to identify harmful content. Although rule-based approaches can detect explicit abusive language, they often fail to recognize implicit toxicity, sarcasm, contextual insults, or variations in spelling and language usage. Therefore, these methods are limited in

their ability to accurately detect complex forms of toxic communication [3], [12].

With the advancement of Natural Language Processing (NLP) and machine learning technologies, more sophisticated approaches have been developed for analysing textual data and identifying harmful content in online platforms. Machine learning algorithms can learn patterns from large datasets and classify comments based on contextual and semantic information. Several traditional machine learning models such as Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, and Random Forest have been widely applied for toxic comment classification tasks [7], [19].

In recent years, deep learning techniques have further improved the performance of toxic comment detection systems. Neural network architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and transformer-based models like BERT can capture complex linguistic patterns and contextual relationships within textual data. These models automatically learn meaningful representations from raw text and often achieve higher classification accuracy compared to traditional machine learning approaches [5], [15]. This research focuses on developing an intelligent toxic comment detection system using machine learning and natural language processing techniques. The proposed system analyses user-generated comments and classifies them into toxic and non-toxic categories based on their linguistic characteristics. The dataset used in this study contains labelled comments representing various types of toxic behaviour, including insults, threats, hate speech, and offensive language.

Several machine learning and deep learning models are implemented and evaluated to determine the most effective approach for toxic comment classification. The performance of these models is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The primary objective of this study is to develop an automated system capable of accurately identifying harmful content and assisting online platforms in maintaining safer and more respectful digital communication environments.

II. LITERATURE SURVEY

The rapid growth of social media platforms and online discussion forums has resulted in an enormous amount of user-generated textual content. While these platforms facilitate communication and information sharing, they have also created environments where harmful content such as hate speech, abusive language, cyberbullying, and offensive comments can spread easily. As a result, researchers have developed various automated approaches to detect toxic comments and maintain safer online communities.

Early studies on toxic comment detection primarily relied on rule-based and lexicon-based approaches. These methods used predefined lists of offensive words and manually constructed rules to identify harmful content. Although these techniques were effective in detecting explicit abusive language, they often failed to recognize implicit toxicity, sarcasm, and context-dependent expressions. Additionally, users could easily bypass such filters by modifying spellings or using slang expressions, which limited the effectiveness of rule-based systems [3].

To overcome these limitations, researchers introduced machine learning-based approaches for toxic comment classification. Machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and Random Forest have been widely applied to analyze textual data and classify comments based on learned patterns. These models typically utilize feature extraction techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) and n-gram representations to convert textual information into numerical feature vectors suitable for machine learning algorithms [12], [19].

Several studies have demonstrated that machine learning models can significantly improve the accuracy of toxic comment detection compared to rule-based systems. However, these models often rely heavily on manual feature engineering and may struggle to capture complex semantic relationships within textual data. As a result, they may produce incorrect classifications when dealing with context-sensitive language or subtle forms of toxicity.

With the advancement of artificial intelligence, deep learning techniques have been increasingly adopted for toxic comment detection tasks. Neural network architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) models are capable of learning complex linguistic patterns directly from textual data without requiring extensive manual feature engineering. These models can capture contextual relationships within sentences and sequential dependencies between words, leading to improved classification performance [13], [14].

More recently, researchers have explored the use of transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) and its variants, for toxic comment classification. Transformer-based architectures generate contextual word embeddings that enable

models to understand the semantic meaning of words within different contexts. Studies have shown that BERT-based models significantly improve classification performance in various natural language processing tasks, including hate speech detection and toxic comment identification [5], [15].

In addition to individual machine learning models, several studies have proposed hybrid and ensemble approaches that combine multiple machine learning and deep learning algorithms to improve prediction accuracy. These approaches leverage the strengths of different models and help reduce classification errors by aggregating predictions from multiple classifiers [8], [17].

Despite the progress made in toxic comment detection, several challenges still remain. Online text often contains informal language, abbreviations, slang, and sarcasm, which can make accurate classification difficult. Furthermore, ensuring fairness and minimizing bias in automated moderation systems remains an important research concern.

Therefore, this study focuses on implementing and comparing multiple machine learning and deep learning models for toxic comment classification. By evaluating their performance on a labeled dataset containing various types of toxic content, the research aims to identify an effective approach for detecting harmful online comments and improving automated content moderation systems.

III. SYSTEM ANALYSIS

A. Existing System

Early approaches for detecting toxic comments on online platforms mainly relied on rule-based filtering systems and keyword matching techniques. These systems used predefined dictionaries containing offensive or abusive words to identify toxic content. When a comment contained any of these words, it was automatically flagged as harmful. Although such methods were able to detect explicit abusive language, they were unable to recognize more complex linguistic patterns such as sarcasm, contextual insults, or indirect toxic expressions.

With the advancement of machine learning technologies, researchers started using supervised learning algorithms to classify online comments. Traditional machine learning models such as Naïve Bayes, Decision Trees, Support Vector Machines, Logistic Regression, and Random Forest have been widely used for toxic comment detection. These models usually analyze text features taken from comments using techniques like bag-of-words, n-grams, and Term Frequency–Inverse Document Frequency (TF-IDF).

Along with traditional machine learning methods, many studies have also used deep learning techniques for toxic comment detection. Neural network models such as Recurrent Neural Networks, Long Short-Term Memory networks, and Convolutional Neural Networks have shown good results in analyzing textual data. These models can learn complex language patterns and understand relationships between words in a sentence, which helps improve classification performance. Even with these improvements, many systems still face challenges when used in real online communication platforms. Online comments often include informal language, slang words, spelling mistakes, emojis, and different contextual meanings. Because of this, it becomes difficult for detection systems to classify comments accurately. As a result, many traditional toxic comment detection systems struggle to provide reliable performance on large social media platforms.

Disadvantages Of The Existing System

Limited contextual understanding: Rule-based systems and simple machine learning models often fail to understand contextual meanings in sentences, which can lead to incorrect classification of comments.

Difficulty in detecting implicit toxicity: Many toxic comments contain sarcasm, hidden insults, or indirect offensive language that traditional models struggle to identify.

Overfitting and underfitting issues: Machine learning models may either memorize training data or fail to capture important linguistic patterns, which reduces prediction accuracy.

High computational requirements: Deep learning models require significant computational resources and large datasets for effective training.

Sensitivity to noisy data: Online comments frequently contain slang, abbreviations, emojis, and spelling variations, which can negatively affect model performance.

Limited scalability: As the volume of user-generated content increases, traditional systems may struggle to efficiently process large datasets.

Lack of adaptability: Existing models may not easily adapt to new types of toxic language, evolving slang, or emerging communication patterns.

B. Proposed System

To address the limitations of existing approaches, this research proposes an intelligent toxic comment detection framework based on machine learning and Natural Language Processing (NLP) techniques. The proposed system automatically analyzes user-generated comments and classifies them into toxic and non-toxic categories.

The system utilizes a labeled dataset containing online comments that represent different forms of harmful language, including insults, threats, hate speech, offensive language, and identity-based attacks. Before model training, the dataset undergoes several text preprocessing steps to improve data quality and prepare the text for machine learning analysis.

The preprocessing stage includes operations such as tokenization, removal of special characters, stop-word removal, text normalization, and lemmatization. These steps help clean the textual data and convert it into a structured format suitable for feature extraction.

After preprocessing, the textual data is transformed into numerical feature vectors using TF-IDF vectorization and word embedding techniques. These representations capture the importance and contextual meaning of words within the dataset.

Multiple machine learning and deep learning models are then trained using the processed dataset. The algorithms evaluated in this study include Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, and Convolutional Neural Networks (CNN). Each model learns patterns associated with toxic and non-toxic comments based on the extracted textual features.

The dataset is divided into training and testing subsets, where approximately 70% of the data is used for model training and 30% for testing. The trained models are evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score to measure their effectiveness in detecting toxic comments.

By integrating advanced NLP techniques with machine learning models, the proposed system aims to provide accurate, scalable, and automated toxic comment detection. This framework can assist social media platforms and online communities in identifying harmful content and maintaining safer digital communication environments.

IV. SYSTEM DESIGN

System Architecture

Below diagram depicts the whole system architecture.

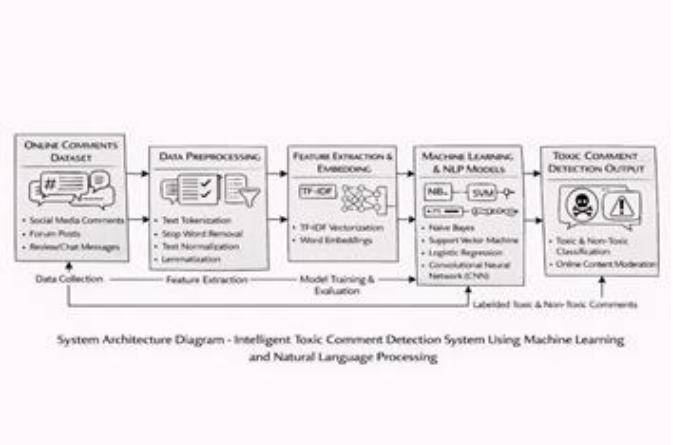

Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION

Modules

This section describes the implementation modules of the proposed toxic comment detection framework. The system follows a structured pipeline consisting of data collection, text preprocessing, feature extraction, machine learning model training, toxic comment detection, and performance evaluation. This modular design improves system scalability, efficiency, and classification accuracy for automated content moderation systems.

A. Data Collection and Preprocessing Module

The first module of the system focuses on collecting and preparing textual data used for toxic comment detection. The dataset contains user-generated comments collected from online platforms such as social media networks, discussion forums, and messaging applications. Each comment in the dataset is labeled as toxic or non-toxic depending on whether it contains harmful content such as insults, threats, hate speech, or offensive language.

Before training the machine learning models, several text preprocessing techniques are applied to clean and standardize the dataset. These preprocessing steps include tokenization, stop-word removal, text normalization, removal of special characters, and lemmatization. These steps help remove unwanted noise from the dataset and convert raw text data into a structured format that can be used for natural language processing and machine learning algorithms.

B. Feature Extraction and Feature Engineering Module
After preprocessing, the textual data is converted into numerical representations using feature extraction techniques. One of the commonly used methods is Term Frequency–Inverse Document Frequency (TF-IDF), which measures the importance of words in a document relative to the entire dataset.

Feature engineering techniques are also applied to identify linguistic patterns associated with toxic language. These methods help the system focus on relevant textual features that contribute to accurate classification. Effective feature selection improves model efficiency, reduces computational complexity, and enhances prediction accuracy.

C. Machine Learning Training Module
In this module, the processed dataset is divided into training and testing subsets, where approximately 70% of the data is used for model training and 30% for model testing.

Multiple machine learning and deep learning algorithms are implemented to classify comments into toxic and non-toxic categories. The algorithms evaluated in this study include:

- Naïve Bayes
- Support Vector Machines (SVM)
- Logistic Regression
- Convolutional Neural Networks (CNN)

Each model is trained using the processed dataset to learn patterns associated with harmful and non-harmful comments. During training, model parameters and hyperparameters are optimized to improve classification performance and reduce prediction errors.

D. Toxic Comment Detection Module
After the training phase, the system is capable of automatically detecting toxic comments from new input data. When a user-generated comment is entered into the system, it undergoes the same preprocessing and feature extraction steps before being analysed by the trained model.

The trained model then classifies the comment as toxic or non-toxic based on the linguistic patterns learned during training. This module can be integrated into online platforms to automatically filter harmful comments and assist moderators in maintaining safe and respectful communication environments.

E. Model Evaluation and Performance Monitoring Module
The performance of the trained models is evaluated using several standard classification metrics, including:

- Accuracy
- Precision
- Recall
- F1-Score

These metrics provide a comprehensive evaluation of the effectiveness of each algorithm in detecting toxic comments. Cross-validation techniques are also used to ensure the reliability and generalization capability of the trained models. Continuous monitoring of model performance is important because online language patterns evolve over time. Updating the system with new training data enables the model to adapt to emerging forms of toxic language and maintain high classification accuracy.

VI. RESULTS AND DISCUSSION

This section presents the experimental results and performance evaluation of the proposed toxic comment detection system using machine learning and deep learning techniques. Several classification algorithms were trained and evaluated using the prepared dataset of labeled online comments. The evaluation focuses on comparing model performance, analyzing classification accuracy, and identifying important textual features that contribute to toxic comment detection.

A. Accuracy Comparison of Machine Learning Models
Multiple machine learning and deep learning algorithms were evaluated to determine the most effective approach for toxic comment classification. The evaluated models include Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), and Convolutional Neural Network (CNN).

Model performance was assessed using evaluation metrics such as accuracy, precision, recall, and F1-score.

Table 1. Performance Comparison of Toxic Comment

Model	Accuracy (%)	Precision	Recall	F1-Score
Naïve Bayes	84.7	0.83	0.82	0.82
Logistic Regression	88.5	0.87	0.86	0.86
Support Vector Machine	90.3	0.89	0.88	0.88
Convolutional Neural Network (CNN)	93.6	0.92	0.91	0.91

From the experimental results, the Convolutional Neural Network (CNN) achieved the highest classification accuracy of 93.6%, outperforming traditional machine learning algorithms. This improved performance can be attributed to the CNN model's ability to capture contextual relationships between words and identify complex linguistic patterns in textual data.

B. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve is used to evaluate the classification performance of the toxic comment detection models by analyzing the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at different classification thresholds.

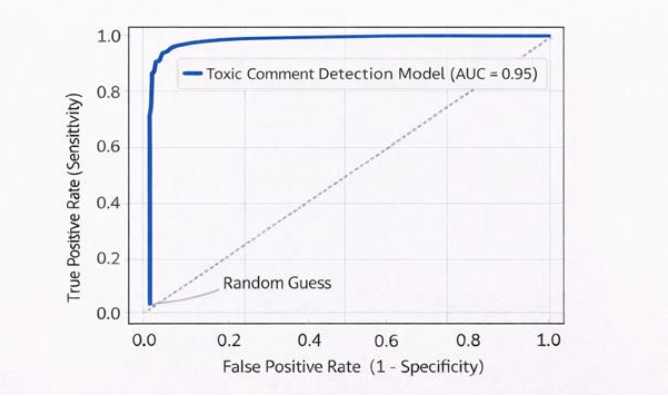

Fig 2. ROC Curve for Toxic Comment Detection Model

The ROC curve shows that the CNN-based model achieved a ROC–AUC score of approximately 0.95, indicating strong classification capability. A ROC curve that approaches the top-left corner of the graph suggests that the model can effectively distinguish between toxic and non-toxic comments with a low false positive rate.

The ROC analysis demonstrates that deep learning models provide reliable performance in detecting harmful online content while maintaining balanced precision and recall values.

C. Text Feature Importance Analysis

To understand the contribution of textual features in detecting toxic comments, a feature importance analysis was conducted using TF-IDF-based word representations.

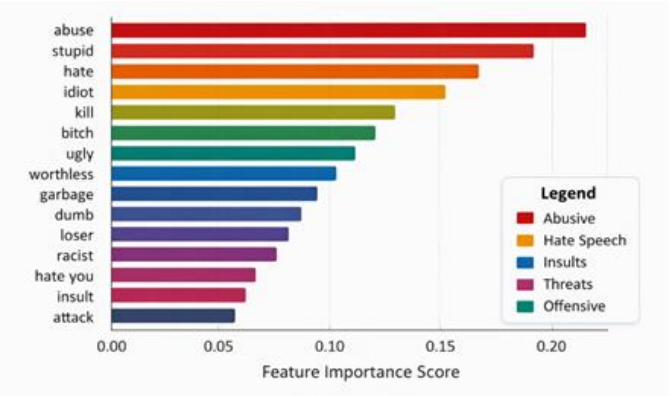

Fig 3. Important Textual Features for Toxic Comment Detection

The analysis revealed that certain words and phrases commonly associated with abusive language, insults, threats, and hate speech had a higher influence on the classification results. These features play a significant role in helping the model distinguish between toxic and non-toxic comments.

Feature importance analysis improves the interpretability of the classification system by highlighting the most influential textual patterns that contribute to toxic language detection. This information can help improve automated moderation systems and assist platform moderators in identifying harmful online interactions.

Overall, the experimental results confirm that combining Natural Language Processing techniques with machine learning and deep learning models significantly improves the accuracy and effectiveness of toxic comment detection systems for online platforms.

VII. CONCLUSION AND FUTURE WORK

This study presented a machine learning–based framework for detecting toxic comments in online communication platforms. The system utilizes Natural Language Processing (NLP) techniques to analyze textual data and classify comments as

toxic or non-toxic. Several machine learning and deep learning algorithms were implemented and evaluated to determine the most effective model for toxic comment classification. Experimental results showed that the Convolutional Neural Network (CNN) model achieved the highest accuracy and demonstrated better performance in identifying complex patterns of harmful language.

The proposed system can assist online platforms in automatically identifying harmful content, reducing cyberbullying, and maintaining safer digital communication environments. By automating the process of toxic comment detection, the system can significantly support moderators in managing large volumes of user-generated content.

Future research may focus on integrating advanced transformer-based models such as BERT and RoBERTa to further improve classification accuracy [5], [15]. Additionally, incorporating multilingual toxic comment detection systems can enable the model to analyze harmful content in multiple languages. Furthermore, combining sentiment analysis and contextual understanding techniques may enhance the system's ability to detect implicit toxicity, sarcasm, and evolving online language patterns.

REFERENCES :

1. J. Risch and R. Krestel, “Toxic comment detection in online discussions,” in Deep Learning-Based Approaches for Sentiment Analysis, 2020, pp. 85–109.
2. S. Kumar and N. Shah, “False information on web and social media: A survey,” arXiv preprint, arXiv:1804.08559, 2018.
3. T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language,” in Proc. International AAAI Conference on Web and Social Media, vol. 11, 2017, pp. 512–515.
4. K. Kurita, A. Belova, and A. Anastasopoulos, “Towards robust toxic content classification,” arXiv preprint, arXiv:1912.06872, 2019.
5. M. Mozafari, R. Farahbakhsh, and N. Crespi, “A BERT-based transfer learning approach for hate speech detection in online social media,” in Proc. International Conference on Complex Networks and Their Applications, Springer, 2020, pp. 928–940.
6. M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Predicting the type and target of offensive posts in social media,” arXiv preprint, arXiv:1902.09666, 2019.
7. D. Patel, P. K. D. Pramanik, C. Suryawanshi, and P. Pareek, “Detecting toxic comments on social media: An extensive evaluation of machine learning techniques,” Journal of Computational Social Science, vol. 8, no. 1, pp. 1–18, 2025.
8. A. Bonetti, M. Martínez-Sober, J. C. Torres, J. M. Vega, S. Pellerin, and J. Vila-Frances, “Comparison between machine learning and deep learning approaches for the detection of toxic comments on social networks,” Applied Sciences, vol. 13, no. 10, p. 6038, 2023.
9. D. Q. Nguyen, T. Vu, and A. T. Nguyen, “BERTweet: A pre-trained language model for English tweets,” arXiv preprint, arXiv:2005.10200, 2020.
10. J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
11. A. Singh, D. Sharma, and V. K. Singh, “Misogynistic attitude detection in YouTube comments and replies: A high-quality dataset and algorithmic models,” Computer Speech & Language, vol. 89, p. 101682, 2025.
12. H. Kajla, J. Hooda, G. Saini, et al., “Classification of online toxic comments using machine learning algorithms,” in Proc. International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2020, pp. 1119–1123.
13. Z. Zhang, D. Robinson, and J. Tepper, “Detecting hate speech on Twitter using a convolution-GRU based deep neural network,” in European Semantic Web Conference, Springer, 2018, pp. 745–760.
14. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in Proc. 26th International Conference on World Wide Web Companion, 2017, pp. 759–760.
15. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, vol. 1, 2019, pp. 4171–4186.
16. S. Carta, A. Corriga, R. Mulas, D. R. Recupero, and R. Saia, “A supervised multi-class multi-label word embeddings approach for toxic comment classification,” in Proc. International Conference on Knowledge Discovery and Information Retrieval (KDIR), 2019, pp. 105–112.
17. V. Maslej-Krešňáková, M. Sarnovský, P. Butka, and K. Machová, “Comparison of deep learning models and various text preprocessing techniques for toxic

comment classification," Applied Sciences, vol. 10, no. 23, p. 8631, 2020.
18. P. Ozoh, A. A. Adigun, and M. Olayiwola, "Identification and classification of toxic comments on social media using machine learning techniques," International Journal of Research and Innovation in Applied Science, vol. 4, no. 11, pp. 142–147, 2019.
19. K. Poojitha, A. S. Charish, M. Reddy, and S. Ayyasamy, "Classification of social media toxic comments using machine learning models," arXiv preprint, arXiv:2304.06934, 2023.
20. M. W. Al Nabki, E. Fidalgo, E. Alegre, and I. De Paz, "Classifying illegal activities on Tor network based on web textual contents," in Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 35–43.