International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Toxic Comment Classification

[1]G Vikas, [2]G Ravi Vardhan, [3]G Somesh, [4]Mr. Lalu Banothu
[1,2,3] UG Scholars, [4] Associate Professor
[1,2,3,4] Department of computer science Engineering,
[1,2,3,4] Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India.

ABSTRACT:

Online abuse and harassment are escalating challenges in the cyber community, requiring robust and efficient content moderation strategies. This study investigates two approaches for identifying toxic comments: training individual classifiers for each facet of toxicity and treating the task as a multi-label classification problem. The facets of toxicity include behaviors like threats, insults, and hate speech, often exhibited simultaneously in a single comment. Using a Kaggle dataset and 10-fold cross-validation, machine learning models such as logistic regression, Naive Bayes, and decision trees are evaluated for performance. A novel preprocessing technique is proposed, transforming multi-label into multi-class classification, which significantly enhances model accuracy. Logistic regression emerges as the most effective model, outperforming others in both binary and multi-class classification tasks. These findings suggest the potential of this preprocessing strategy to boost the performance of more sophisticated neural network-based models, offering a scalable solution for real-world toxic content moderation systems.

1 INTRODUCTION:

The widespread use of social media and the internet has significantly transformed modern life, enabling people to share opinions, critiques, and discussions through comments. These platforms have become a vital medium for communication, but they also present challenges due to the unregulated and often hostile nature of online interactions. Text classification, a core technique in applications such as sentiment analysis, topic categorization, and information filtering, plays a crucial role in addressing these issues. However, online comments are often noisy, brief, and less structured, making it difficult to analyze them effectively.

The growing prevalence of online abuse, harassment, and cyberbullying highlights the need for robust methods to identify harmful content. Studies, such as one conducted by the Pew Research Center in 2017, indicate that 41% of American internet users have experienced harassment, and 61% have witnessed such behavior. These findings emphasize the urgent need for content moderation to protect users and foster a safer online environment. Many platforms have policies to address harassment, such as blocking accounts or removing harmful comments, but these policies rely on accurately identifying toxic content.

Online comments can express negativity in various forms, such as distrust, abusive language, or derogatory remarks. Analyzing and classifying these comments based on toxicity levels is crucial for implementing moderation policies effectively. This study aims to develop a methodology for identifying and classifying online comments according to their toxicity, enabling platforms to enforce rules and reduce harmful interactions. By filtering toxic content, this work contributes to creating healthier and more constructive online discussions.

1.1 OBJECTIVE:

The objective of this project is to develop a method for detecting and mitigating online toxicity in user comments on social media platforms. With the increasing volume of content generated daily, toxic

comments are negatively affecting the quality of online interactions, limiting healthy discussion and preventing individuals from expressing dissenting opinions. This project aims to identify antisocial behavior in online forums, offering a reliable solution for content moderation. While previous efforts, such as crowd-sourcing and comment reporting systems, have had limited success, this study seeks to employ a more effective technique to detect and address toxicity in user-generated content.

1.2 SCOPE OF THE WORK:

The scope of this project involves developing a system to detect and classify toxic comments on online platforms. It includes data collection from publicly available datasets like Kaggle's Toxic Comment Classification dataset, followed by preprocessing tasks such as text cleaning, tokenization, and feature transformation using techniques like TF-IDF or word embeddings. Various machine learning models, including Logistic Regression, Naive Bayes, and Decision Trees, will be implemented for both binary and multi-label classification of toxicity. Model performance will be evaluated using metrics such as accuracy, precision, and recall, with a focus on identifying different levels and types of toxicity. The project aims to create a robust solution that can be applied to real-world moderation systems, reducing cyberbullying and promoting healthier online discussions. Future work may include the integration of more advanced models and personalization features for improved accuracy.

1.3 PROBLEM STATEMENT:

The increasing prevalence of toxic comments on social media platforms is undermining healthy online discourse and creating a toxic environment that discourages free expression and respectful debate. Despite efforts such as crowd-sourcing moderation and comment reporting systems, current methods are often inadequate in effectively detecting and addressing toxic content in real time. This leads to the persistence of online harassment, cyberbullying, and the suppression of dissenting opinions. Therefore, there is an urgent need for a more robust and automated system to accurately identify and classify toxic comments based on various forms of negativity, such as insults, threats, and hate speech. This project seeks to develop an effective solution for detecting online toxicity using advanced machine learning techniques to improve content moderation and create safer online spaces for users..

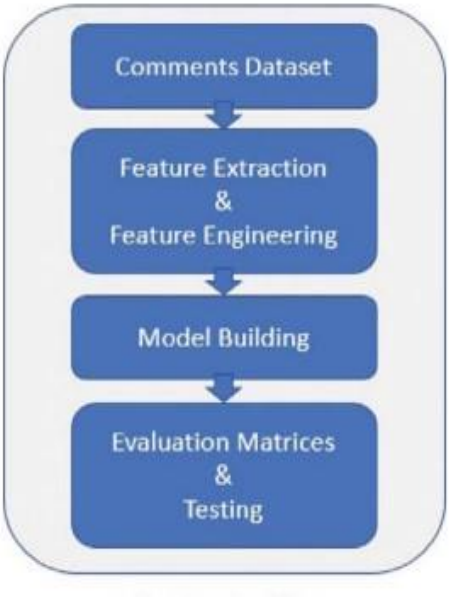
1.4 EXISTING SYSTEM:

In recent years, Convolutional Neural Networks (CNNs) have gained popularity in text classification and natural language processing tasks. These models work by using both distributed and discrete word embeddings, without needing explicit syntactic or semantic knowledge of the language. An empirical study by Zhang et al. demonstrates that CNNs utilizing character-level features can be an effective approach for text classification. Additionally, a recurrent CNN model has been proposed, which outperforms traditional CNN models and other well-established classifiers by capturing contextual information through its recurrent structure. This model constructs text representations using CNNs, while effectively handling word order for categorization. A more recent development involves a CNN-based model that leverages word embeddings to encode text, incorporating semantic, sentiment, and lexicon embeddings. This approach integrates three different attention mechanisms—attention vector, LSTM (Long Short-Term Memory) attention, and attentive pooling—into the CNN architecture. To further enhance performance, techniques such as Cross-modality Consistent Regression (CCR) and transfer learning have been introduced. Notably, this is the first application of CCR and transfer learning to textual sentiment analysis. Experiments on two distinct datasets show that the proposed CNN attention models achieve superior or near-state-of-the-art results compared to existing models.

1.4.1 Existing System Disadvantages:

One of the main disadvantages of the existing system is its unexplained functioning, where the underlying processes or reasons for its behavior are not clearly understood. This lack of transparency makes it difficult to troubleshoot or improve the system effectively. Additionally, the duration for which the network operates or the time it takes to complete tasks remains uncertain. This unpredictability can lead to inefficiencies, making it challenging to optimize the system for real-time applications or large-scale deployments.

1.5 SYSTEM ARCHITECTURE:



1.5.1 EXPLANATION:

This project aims to tackle the growing issue of online toxicity by classifying toxic comments in social media and online discussion forums using machine learning techniques. The dataset used is sourced from Kaggle, containing comments labeled as toxic or non-toxic. The project explores two classification approaches: binary classification for each type of toxicity (such as threats, insults, and negativity) and multi-label classification to identify multiple toxic aspects in a single comment.

Data preprocessing steps, including handling missing values and encoding categorical features, are applied to prepare the data for training. Various machine learning algorithms, such as Logistic Regression, Naive Bayes, and Decision Trees, are implemented and evaluated using 10-fold cross-validation for model robustness. Among all tested models, Logistic Regression outperforms others, providing the best accuracy in classifying toxic comments. The findings suggest that this approach can help automate the detection of harmful content, contributing to a safer online environment.

1.6 PROPOSED SYSTEM

In this paper, we propose a simulation model to evaluate the effectiveness of the suggested algorithm by generating nighttime images with different levels of contrast and noise. The algorithm is designed to process a variety of images efficiently, avoiding common issues like ghosting and halo artifacts. To assess its performance, both full-reference and blind performance metrics are used, and the results show that the proposed method provides state-of-the-art performance in terms of both objective measurements and visual quality, outperforming existing methods. The evaluation process of these algorithms is essential to understand their effectiveness. Currently, algorithm evaluation is carried out from two main perspectives: qualitative and quantitative. Image Quality Assessment (IQA) plays a key role in image processing, contributing to enhancements and denoising techniques. Existing IQA methods typically address issues like compression, transmission errors, noise, and blurring artifacts. Depending on the availability of reference images, IQA approaches are categorized into three groups: full-reference (FR), reduced-reference (RR), and no-reference (NR)/blind metrics.

1.6.1 PROPOSED SYSTEM ADVANTAGES:

The proposed system significantly improves the contrast and effectively eliminates noise from the data. It is highly efficient, delivering strong results with minimal resource consumption. Additionally, the system can be implemented quickly, ensuring a fast and practical solution for real-world applications.

2 DESCRIPTION:

2.1 GENERAL:

This project aims to classify toxic comments on online platforms using machine learning techniques. The dataset, sourced from Kaggle, contains various features such as comment content and user information. Two approaches are used: binary classification for detecting specific types of toxicity (e.g., insults, threats) and multi-label classification for identifying multiple toxic aspects in a comment.Several machine learning algorithms, including Logistic Regression, Naive Bayes, and Decision Trees, are tested. A novel preprocessing technique transforms multi-label classification into a multi-class problem, improving model performance. After applying 10-fold cross-validation, Logistic Regression outperforms other models in accuracy and efficiency.The results demonstrate that the approach can effectively detect toxic behavior, providing a foundation for creating safer online environments by identifying and penalizing offenders.

2.2 METHODOLOGIES

2.2.1 MODULES NAME:

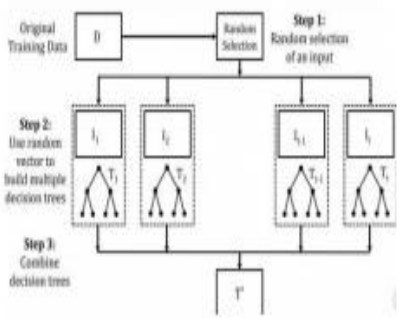
1. Database
2. Feature Study and Selection
3. Data visualization
4. Data Processing
5. Learning Algorithm

1. Dataset:

The dataset used in this study is sourced from the UCI Machine Learning Repository [8], originally derived from the 1994 U.S. Census database by Barry Becker. It contains 48,842 records with 14 attributes, including both categorical (e.g., age, education, occupation) and continuous (e.g., weekly working hours, capital gain/loss) variables. The goal is to predict whether an individual earns more than $50,000 per year based on these attributes.

2. Feature Study and Selection:

Based on the feature importance scores from the Extra Trees Classifier (Table 1), the most significant features were chosen for the model. A visual overview of the Extra Trees Classifier (also known as Extremely Randomized Trees) is provided in Fig. 1. As a result, Features F9 (race) and F14 (native-country) were excluded due to their minimal importance scores.



3. Data Visualization:

To gain a better understanding of the central tendencies of the continuous features, data visualization was carried out using Box and Whisker plots. These plots provide a clear representation of the distribution, highlighting the median, quartiles, and potential outliers within the data.

4. Data Processing:

Before processing the Adult Dataset, several preprocessing techniques were applied. Missing values in categorical features such as work class, occupation, and native-country were addressed by marking them with a placeholder ('?') and assigning a unique category to prevent data loss. For encoding categorical features, two methods were used: label

encoding, where each category was assigned a numeric value starting from 0, and one-hot encoding, which was applied to categorical features with more than two categories, representing each category as a binary value. The sex attribute, having only two categories, was left in its binary form to avoid dimensionality issues. Additionally, the dataset was shuffled consistently to ensure that all attribute categories were well-represented in both the training and validation sets. Finally, the data was split into training (80%) and testing (20%) sets to prepare for model evaluation.

5. Learning Algorithm:

The predictive model in this study is built using the Gradient Boosting Classifier (GBC), which is a powerful ensemble learning technique known as Boosting. Boosting is an iterative process where multiple weak learners, typically decision trees, are trained sequentially. In contrast to traditional methods where models are built independently, boosting builds decision trees one after another. Each new tree attempts to correct the errors made by the previous trees. Specifically, when training the Gradient Boosting Classifier, the algorithm focuses on the mistakes made by earlier trees, and in the subsequent iterations, it tries to reduce these errors, thereby improving the accuracy of the model step by step. This allows the model to become more accurate and efficient at fitting the training data as more trees are added to the sequence. The result is a strong predictive model that performs well even with complex data sets.

2.3 TECHNIQUE USED OR ALGORITHM USED

2.3.1 EXISTING TECHNIQUE:

Convolutional Neural Networks(CNN):

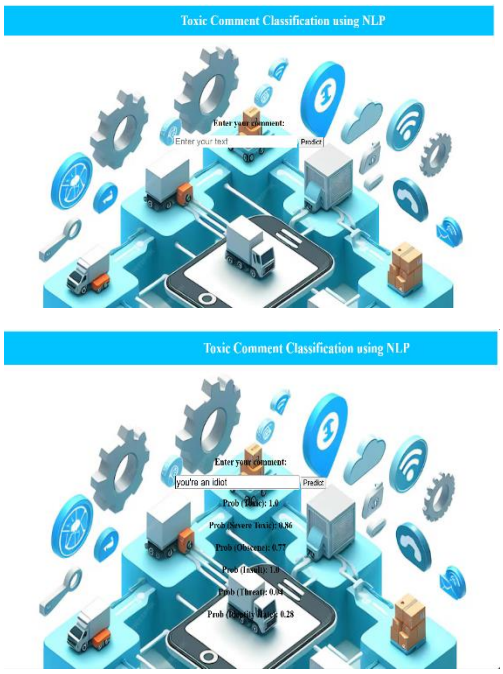
The techniques used in the discussed content involve several advanced methods for text classification and natural language processing (NLP). Convolutional Neural Networks (CNNs) are applied to process both distributed and discrete word embeddings without relying on syntactic or semantic knowledge of language. Specifically, character-level CNNs are utilized for text classification, allowing the model to work directly with raw text characters, making it effective without predefined features. Additionally, a recurrent CNN model is introduced, combining the benefits of recurrent neural networks (RNNs) to capture contextual information in text. To further enhance text representation, various word embeddings, including semantic, sentiment, and lexicon embeddings, are employed. The model also integrates different attention mechanisms, such as attention vectors, LSTM attention, and attentive pooling, to focus on relevant parts of the input text. Moreover, Cross-Modality Consistent Regression (CCR) is used to improve the relationship between different modalities, and transfer learning is applied to leverage pre-trained models, boosting performance in text classification and sentiment analysis tasks. These combined techniques lead to a powerful and efficient approach for text categorization and sentiment analysis.

2.3.2 PROPOSED TECHNIQUE USED OR ALGORITHM USED:

The techniques employed in the above content focus on evaluating and improving image quality. A simulation model is used to generate nighttime images with varying levels of contrast and noise to test the algorithm's performance. The algorithm is designed to process these images without introducing ghosting or halo artifacts, ensuring high-quality output. To assess the effectiveness of the algorithm, both full-reference and blind performance metrics are utilized. Full-reference metrics compare the processed image with a reference image, while blind metrics evaluate the image independently, without needing a reference. Image Quality Assessment (IQA) techniques play a crucial role in enhancing and denoising images, addressing challenges such as noise and compression artifacts. Additionally, IQA approaches are categorized into three types: full-reference, reduced-reference, and no-reference (blind) metrics, depending on the availability of reference images for comparison. These techniques collectively contribute to evaluating and enhancing the image quality, ensuring the

proposed algorithm performs effectively compared to existing methods.

3. RESULT:



In this project, we aimed to classify toxic comments using machine learning techniques. The dataset from Kaggle was preprocessed, and two classification approaches were applied: binary classification for each toxicity facet and multi-label classification. Among the models tested, Logistic Regression outperformed Naive Bayes and Decision Trees in accuracy and efficiency. A novel preprocessing strategy that transformed the multi-label problem into a multi-class classification improved model performance significantly. Key features like specific words and comment structure were crucial in detecting toxicity. The results suggest Logistic Regression is effective for identifying online harassment and can be applied to more complex models.

4. FUTURE ENHANCEMENT:

In future developments, this research can be extended by utilizing recurrent neural networks (RNNs) to improve model performance. Furthermore, incorporating community-bias analysis could refine the models, allowing them to adjust and align with the toxicity standards specific to various online communities. This enhancement would enable more precise identification of toxic comments, tailored to the unique norms of different groups.

5. CONCLUSION:

The rapid growth of online discussions, particularly with the rise of social media, has led to a significant increase in cyber harassment. In response, many social platforms have established policies to penalize offenders. However, identifying the level and type of toxicity or negativity in comments remains a challenging task. To tackle this, the study introduces two approaches. Initially, a dataset from a Kaggle competition is analyzed to identify key features that can help detect toxic comments. Two different classification methods are then applied: binary classification for each type of toxicity and multi-label classification. Additionally, several machine learning techniques are tested for each classification method. The performance of these models is evaluated using 10-fold cross-validation. The results show that Logistic Regression outperforms other methods, including Naive Bayes and Decision Tree Classifier. The promising performance of Logistic Regression, especially after implementing a novel preprocessing strategy, indicates its potential applicability to neural network models.

6 REFERENCES:

[1] G. Song, Y. Ye, X. Du, X. Huang, and S. Bie, “Short Text Classification: A Survey.” Journal of Multimedia, vol. 9, pp.635- 643,2014.

[2] Y. Rui, C. Xian-bin, L. Kai, “Dynamic Assembly Classification Algorithm for Short Text,” ACTA ELECTRONICA SINICA, vol. 37(5), pp. 1019-1024, 2009.

[3] C. Aggarwal and Z. ChengXiang, “A survey of text classification algorithms,” Mining text data, Springer, pp. 163- 222, 2012.

[4] Maeve Duggan, "Online harassment," Pew Research Center, 2014.

[5] "Wikipedia: No personal attacks.," Wikipedia, https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks.

[6] M. Duggan, "Online Harassment 2017," Pew Research Center: Internet, Science & Tech, Jul 11, 2017.

[7] "Harassment consultation 2015-Meta," Available: https://meta.wikimedia.org/wiki/Harassment_consultation_2015. [Accessed: May 28, 2019].

[8] A. Maus, "SVM approach to forum and comment moderation," Cl. Proj. CS, 2009.

[9] A. Mosquera, L. Aouad, S. Grzonkowski, and D. Morss, "On Detecting Messaging Abuse in Short Text Messages using Linguistic and Behavioral patterns," ArXiv Prepr. ArXiv14083934, 2014.

[10] P. Goyal and G. S. Kalra, "Peer-to-peer insult detection in online communities," IITK Unpubl., 2013.

[11] E. Wulczyn, N. Thain, and L. Dixon, "Exmachina: Personal attacks seen at scale," in Proceedings of the 26th International Conference on World Wide Web, pp. 1391-1399, 2017.

[12] L. Song, R. Y. Lau, and C. Yin, "Discriminative Topic Mining for Social Spam Detection.," PACIS, pp. 378, 2014.

[13]"Perspective," Available: https://www.perspectiveapi.com/#/. [Accessed: July 12, 2019]. [14] S. V. Georgakopoulos, S. K. Tasoulis, A. G. Vrahatis, and V. P. Plagianakos, "Convolutional Neural Networks for Toxic Comment Classification," ArXiv180209957 Cs, Feb. 2018.

[15] "sklearn.linear_model.LogisticRegression — scikit-learn 0.19.2 documentation." Available: http://scikitlearn.org/stable/modules/generated/sklearn.linear_model.Logistic Regression.html. [Accessed: May 16, 2018].

[16] "Toxic Comment Classification Challenge," Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classificationchallenge. [Accessed: May 28, 2018].

[17] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., 2011, vol. 12, pp. 2825-2830.

[18] E. Jones, T. Oliphant, P. Peterson, and others, SciPy: Open source scientific tools for Python, 2001.

[19] Bird, Steven, Edward Loper and Ewan Klein , Natural Language Processing with Python. O'Reilly Media Inc, 2009.

[20] J. D. Hunter, "Matplotlib: A 2D graphics environment," Comput. Sci. Eng., 2007, vol. 9(3), pp. 90-95.

[21] MR. Murty, JVR. Murthy, and P. Reddy. "Text Document Classification based-on Least Square Support Vector Machines with Singular Value Decomposition," International Journal o f Computer Applications, vol. 27(7), pp. 21-26, 2011.