Toxic Comment Classification and Mitigation in Social Media Platforms

Sunitha Sabbu, Rajani D., Shaik Jaheed D., Sreeja Y. and Venkata Naga Hemanth Reddy B.
Department of CSE (AI & ML), Srinivasa Ramanujan Institute of Technology, Rotarypuram Village, B K Samudram Mandal, Anantapuramu - 515701, Andhra Pradesh, India

Keywords: Toxic Comments, NLP, Deep Learning, LSTM, Embeddings.

Abstract: Toxic online discussions have become a growing concern on digital platforms, necessitating automated solutions for detecting harmful content. This paper presents a Deep learning-based system that uses NLP methods to classify comments. Comments is categorized into toxic or non-toxic-are recognized by the system. A LSTM neural network-based classifier, integrated with Word2Vec skip-gram embedding and a fully connected network, is employed to process and classify user-generated comments. To facilitate real-time toxicity detection, we integrated the trained model with Twitter clone that allows users to input comments and receive toxicity predictions instantly. The dataset, sourced from a publicly available toxic comment corpus, consists of over 110,480 comments, providing a robust foundation for training. Model evaluation demonstrates high accuracy and Precision, with misclassification challenges observed in sarcasm and implicit toxicity detection. This research contributes to the field of online content moderation, offering an efficient and scalable approach for classifying comments. Future enhancements include advanced deep learning models and embedded context to improve nuanced toxicity identification in internet-based discussions.

1 INTRODUCTION

The exponential rise in digital communication platforms has significantly transformed how people interact, fostering dynamic social exchanges. However, this growth has also led to the proliferation of toxic comments, which can create hostile online environments and discourage meaningful dialogue. The sheer volume of user-generated content makes manual moderation impractical, necessitating automated systems capable of effectively detecting and mitigating toxic language.

Traditional rule-based systems, which rely on keyword matching, struggle to capture the complexity of language, including sarcasm and contextual nuances. Deep learning models, particularly those employing statistical and deep learning approaches, provide a more dynamic and scalable solution. By analyzing patterns in large datasets, these models can effectively classify comments into toxic and non-toxic.

This study presents an automated toxic comment detection system using a LSTM-based Classifier Combined with Word2vec skip-gram embedding to effectively classify toxic comments. The model is trained on a large dataset of over 110,480 comments labeled for toxicity. To enhance usability, we integrated the trained model with twitter clone, allowing users to input comments and receive real-time toxicity predictions.

By integrating real-time toxicity detection with twitter clone, this approach provides a practical solution for moderating online discussions The LSTM approach offers computational efficiency while maintaining high accuracy in toxicity classification.

This paper is organized as follows: Section II discusses related work on toxic comment detection, comparing traditional and modern approaches. Section III outlines the methodology, including data preprocessing, model architecture, and the model deployment. Section IV presents experimental results, discussing accuracy, precision. Section V concludes the study and suggests future enhancements, such as incorporating advance deep learning techniques and expanding the dataset for improved detection of nuanced toxicity.

Sabbu, S., D., R., D., S. J., Y., S. and B., V. N. H. R.
Toxic Comment Classification and Mitigation in Social Media Platforms.
DOI: 10.5220/0013903100004919

In *Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25 2025) - Volume 3*, pages 627-631
ISBN: 978-989-758-777-1

2 RELATED WORKS

Toxic comment detection has been a prominent research area within Natural Language Processing (NLP) and Machine Learning (ML), with various approaches explored to enhance classification accuracy and fairness in automated moderation systems. Early methods relied on rule- based systems and keyword filtering, which struggled to handle contextual variations, sarcasm, and evolving language patterns. Recent advancements have shifted towards ML and deep learning models, which provide more adaptable and robust solutions by learning complex patterns from large datasets.

Several studies have examined the impact of social networks on user behavior and the need for automated moderation systems to address toxic interactions. For instance, Arab and Díaz investigated the influence of social platforms on adolescent behavior, emphasizing the importance of automated systems in minimizing harmful content. Risch and Krestel demonstrated the effectiveness of deep learning models for sentiment analysis and toxic comment detection by leveraging neural networks to capture contextual nuances. Meanwhile, Modha et al. highlighted the challenges of hate speech detection in Indo-European languages, underscoring the limitations of traditional classifiers in multilingual settings.

Other research efforts have explored sentiment-aware models and algorithmic efficiency. Brassard-Gourdeau and Khoury integrated sentiment analysis into toxicity classification, improving detection accuracy by considering emotional context. In contrast, Kiilu et al. utilized Naïve Bayes algorithms for hate speech detection on Twitter, demonstrating computational efficiency in low-resource environments. Miró-Llinares et al. developed a specialized algorithm for detecting hate speech in digital microenvironments, highlighting the importance of domain-specific adaptation in toxicity classification.

Multilingual toxicity detection has also been a critical focus, as cultural and linguistic variations influence how toxic comments are expressed. Almatarneh et al. compared multiple supervised classifiers for identifying hate speech in English and Spanish tweets, showcasing the challenges of multilingual model development. Davidson et al.examined bias issues in automated hate speech detection models, advocating for fairness-aware NLP techniques to minimize algorithmic discrimination. Venkit and Wilson further analyzed biases against people with disabilities, emphasizing the need for inclusive and equitable toxicity detection systems.

Recent innovations in feature engineering and model architecture have also contributed to improved toxicity classification. Shtovba et al. introduced syntactic dependency-based approaches that leverage grammatical structures for more accurate classification. Qader et al. analyzed Bag of Words (BoW) methods, discussing their applications and limitations in text classification. Li et al. reviewed advancements in feature selection, with a focus on TF-IDF-based models for efficient text representation. Robertson provided foundational insights into Inverse Document Frequency (IDF), explaining its role in enhancing text vectorization and information retrieval.

Building on these prior studies, this research integrates logistic regression with TF-IDF vectorization for efficient real-time toxicity detection. Unlike deep learning models that require extensive computational resources, the proposed system balances accuracy and scalability while leveraging user feedback for continuous model adaptation. This hybrid approach provides a practical solution for moderating online discussions while maintaining high classification accuracy Future studies will examine hybrid models that mix deep learning architectures with conventional machine learning techniques to improve contextual understanding in toxicity detection across diverse online environments.

3 METHODOLOGY

3.1 Dataset

This study utilizes a publicly accessible dataset specifically designed for toxic comment classification. It consists of approximately 110,480 user-generated comments that are labeled as 0(non-toxic), 1(toxic). Since each comment can be associated with either 0 or 1 labels, the classification problem is treated as a binary-label classification task. To ensure the model's effectiveness across various scenarios, the dataset encompasses a diverse range of comments sourced from different online platforms, thus maintaining a balanced distribution of toxic and non-toxic instances.

To enhance the classification model's accuracy, the dataset undergoes comprehensive pre-processing. This involves several crucial steps, including the removal of duplicate entries and the handling of missing values to ensure data integrity. Furthermore, text normalization is performed to create uniformity, involving the elimination of special characters, punctuation marks, and extraneous symbols that do not contribute to the classification task. These pre-processing measures help in standardizing the text and reducing noise, ultimately leading to improved model performance.

3.2 Data Preprocessing

To improve model performance, the following preprocessing steps are applied:
Lowercasing: Converts all text to lowercase to ensure uniformity.

Removing special characters and punctuation: Eliminates non-alphanumeric symbols that do not contribute to classification.

Expanding contractions: Converts common contractions (e.g., "can't" → "cannot") into their full forms.

Stopword removal: Filters out frequently used words that do not add significant meaning.

Tokenization: Splits text into individual words to enhance feature extraction.

Word2Vec Skip-gram Embedding: Converts the processed text into numerical feature vectors for classification.

After completing these pre-processing steps, this stratified split ensures that both subsets maintain the same proportion of toxicity categories, enabling accurate model evaluation and performance assessment.

3.3 Model Architecture

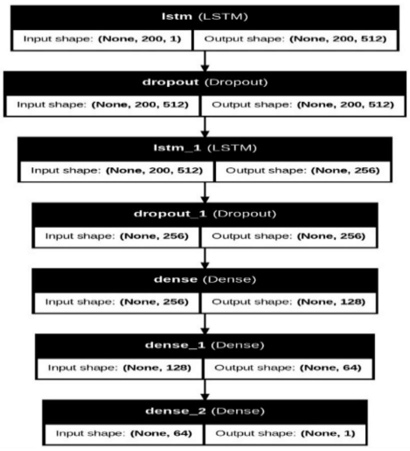


Figure 1: Architecture of deep learning model.

This Figure 1 employs a LSTM classifier with word2vec skip-gram embedding, where each toxicity category is treated as an independent binary classification task. The model architecture is designed to ensure computational efficiency, scalability, and high classification accuracy while maintaining interpretability.

Class Imbalance: There is an imbalance in target class where the non-toxic instances are high compared to toxic instances. To balance this Random under-sampling technique is used. This technique randomly removes the non-toxic instances and makes the target class balance.

As shown in Figure 2, the original dataset exhibits a significant class imbalance, which can negatively impact model performance. However, after applying resampling techniques, the class distribution becomes more balanced, as illustrated in Figure 3.

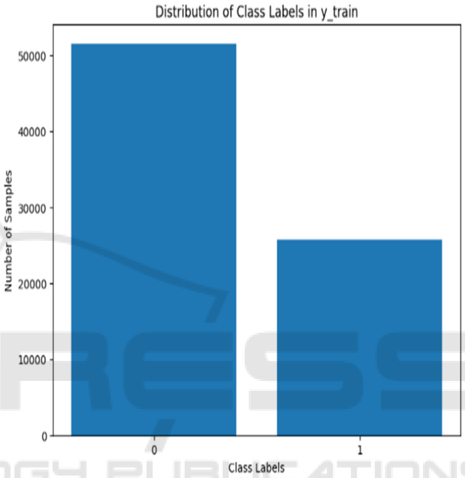


Figure 2: Class imbalance visualization.

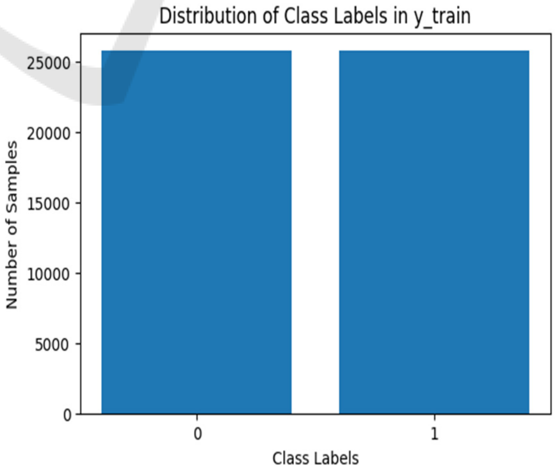


Figure 3: Balanced class distribution after resampling.

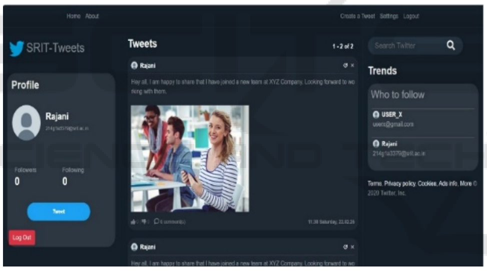
- **Feature Extraction:** Converting text data into numerical feature vectors that reflect the importance of words relative to the dataset.
- **Hyperparameter Tuning:** Hyperparameter Tuning is performed by adjusting the parameters of a model and observing the metrics.

This systematic tuning process enhances the model's performance by identifying the optimal configuration for each classification task. Overall, this architecture balances accuracy and interpretability, making it a suitable choice for real-time toxic comment detection.

3.4 Model Deployments

The trained model is integrated with twitter clone, to facilitate real-time comment classification. The system consists of the following features:

- **User Input Interface:** Allows users to post and submit comments for toxicity analysis.
- **Prediction Engine:** Processes input text and toxicity predictions based on the trained model. Figure 4 Shows the User Interface.s



Hey all, I am happy to share that I have joined a new team at XYZ Company. Looking forward to working with them.

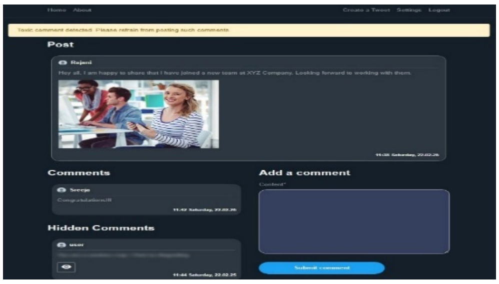
11:38 Saturday, 22.02.25

Terms, Privacy policy, Cookies, Ads info, More © 2020 Twitter, Inc.

Figure 4: User interface.

3.5 Toxic Comment Mitigation

Toxic comments are mitigated by showing alert message and preventing the display to the users for healthier interactions. Real Time Comment Mitigation Shown in Figure 5.



Toxic comment detected. Please refrain from posting such comments.

Hey all, I am happy to share that I have joined a new team at XYZ Company. Looking forward to working with them.

11:38 Saturday, 22.02.25

Congratulations!!!

11:42 Saturday, 22.02.25

11:44 Saturday, 22.02.25

Figure 5: Real time comment mitigation.

4 RESULTS AND ANALYSIS

The performance of the LSTM-based toxicity detection model was evaluated using a stratified test set, ensuring a balanced representation of all toxicity categories. To assess the model's effectiveness, key classification metrics were employed, including accuracy, precision, recall, and F1-score. The model demonstrated high accuracy in distinguishing between toxic and non-toxic comments, correctly classifying the majority of test samples. In terms of precision, the model excelled at identifying explicit toxic content but occasionally misclassified non-toxic comments as toxic, indicating room for improvement in reducing false positives. The recall metric revealed the model's capability to correctly identify a significant proportion of toxic comments, although it faced challenges in recognizing nuanced or context-dependent toxicity. The model continuously performed well in classification, according to the F1-score, which weighs precision and recall. Overall, the evaluation indicates that while the model effectively detects explicit toxicity, it requires further refinement to accurately classify subtle and implicit toxic expressions.

5 DISCUSSION AND FUTURE WORK

5.1 Discussion

The results demonstrate that combining LSTM with Word2Vec skip-gram embedding is an effective approach for comment classification. The system efficiently classifies comments into toxic or non-toxic, enabling automated moderation on online platforms. This approach provides a balance of interpretability and computational efficiency, making it suitable for real-time applications. Additionally, the integration of a twitter clone web interface allows seamless user interaction and real-time toxicity classification.

However, several challenges were identified. The model struggles with implicit toxicity and sarcasm, which require contextual understanding beyond the capabilities of traditional machine learning models, leading to false negatives. These limitations suggest the need for advanced techniques, such as contextual embeddings and deep learning models, to improve nuanced toxicity detection and reduce misclassification rates.

5.2 Future Work

To enhance the performance of the toxicity detection model, several improvements are proposed. Advanced deep learning model such as BERT can be implemented to capture contextual nuances and implicit toxicity more effectively. Hybrid models combining rule-based linguistic features with machine learning algorithms can improve sarcasm detection and contextual understanding. For scalability, the model can be optimized for real-time deployment with cloud-based APIs to support high-throughput applications. Additionally, integrating user feedback through active learning frameworks will enable continuous model adaptation to evolving language patterns. These enhancements will significantly improve the model's accuracy, contextual understanding, and scalability in real-world toxicity detection scenarios.

6 CONCLUSIONS

The increasing prevalence of toxic comments on online platforms necessitates robust automated moderation systems. In order to efficiently identify comments as either harmful or non-toxic, this paper proposes a deep learning-based method for detecting toxic remarks utilizing LSTM and Word2Vec skip-gram feature extraction. The model is integrated with a twitter clone web application, allowing real-time prediction.

According to study findings, the model is highly accurate in identifying toxic language and performs well in detecting toxicity. Despite the promising results, limitations remain, particularly in handling rare toxicity categories and complex language structures. Future improvements include advanced deep learning models for better contextual understanding, data augmentation techniques and hybrid approaches integrating linguistic rules with machine learning models.

Overall, this research advances by presenting an automated content moderation by providing an efficient, scalable, and adaptive framework for real-time toxic comment detection, enhancing the safety and quality of online interactions.

REFERENCES

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet for English Tweets," EMNLP, Assoc. Comput. Linguistics, Cedarville, OH, USA, Nov. 2020.

E. Brassard-Gourdeau and R. Khoury, "Sentiment Information for Toxicity Detection," Proc. Abusive Lang. Online Workshop, Florence, Italy, Aug. 2019.

F. Miró-Llinares, A. Moneva, and M. Esteve, "Algorithm for Hate Speech Detection in Microenvironments," Crime Sci., vol. 7, p. 15, 2018.

J. Risch and R. Krestel, "Toxic Comment Detection in Online Discussions," Deep Learning-Based Sentiment Analysis, Springer, 2020.

K. K. Kiilu, G. Okeyo, R. Rimiru, and K. Ogada, "Naïve Bayes Algorithm for Hate Speech Detection," Int. J. Sci. Res. Publ., vol. 8, pp. 99–107, 2018.

L. E. Arab and G. A. Díaz, "Social Network Impact on Adolescence," Rev. Medica Clin. Condes, vol. 26, pp. 7–13, 2015.

P. Wiemer-Hastings, "Latent Semantic Analysis," Encycl. Lang. Linguist., vol. 2, pp. 1–14, 2004.

S. Robertson, "On Theoretical Arguments for Inverse Document Frequency," J. Doc., vol. 60, pp. 503–520, 2004.

S. Modha, T. Mandl, P. Majumder, and D. Patel, "Hate Speech Identification in Indo-European Languages," FIRE ’19, Kolkata, India, Dec. 2019.

S. Almatarneh, P. Gamallo, F. J. R. Pena, and A. Alexeev, "Classifiers for Hate Speech in English and Spanish Tweets," Digit. Libr. Crossroads, vol. 11853, pp. 25–30, 2019.

S. Shtovba, O. Shtovba, and M. Petrychko, "Toxic Comment Detection Using Syntactic Dependencies," CEUR Workshop 2353, Zaporizhzhia, Ukraine, Apr. 2019.

T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and Offensive Language Classification," arXiv, 2017.

T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Word Representations in Vector Space," arXiv, 2018.

W. A. Qader, M. M. Ameen, and B. I. Ahmed, "Bag of Words Overview: Applications and Challenges," Int. Eng. Conf., Erbil, Iraq, Jun. 2019.

Y. Li, T. Li, and H. Liu, "Advances in Feature Selection and Applications," Knowl. Inf. Syst., vol. 53, pp. 551–577, 2017.