ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 15, Issue, 02, pp. 67737-67742, February, 2025
https://doi.org/10.37118/ijdr.29239.02.2025

REVIEW ARTICLE OPEN ACCESS

MULTILINGUAL TOXIC COMMENTS CLASSIFICATION USING BERT

***A. Akshaya, K. Sindhuja, N. Rohan and Y. Sahas**

Computer Science Business System, B V Raju Institue of Technology Hyderabad, India

ARTICLE INFO

Article History:
Received 10th December, 2024
Received in revised form
16th December, 2024
Accepted 25th January, 2025
Published online 27th February, 2025

Key Words:
Multilingual Toxic comments, Bert (Bidirectional Encoder Representations from Transformers), NLP (Natural language processing).

****Corresponding author: Akshaya, A***

ABSTRACT

The swift expansion of online platforms has led to a surge in toxic comments, disrupting digital communities and adversely affecting users. Tackling this pervasive issue presents significant challenges, particularly in a multilingual context, as most available solutions tend to focus primarily on English. This project presents a multilingual toxic comments classification system harnessing Multilingual BERT (mBERT) capabilities. By utilizing mBERT's proficiency in various languages, the system can proficiently detect and classify toxic content—ranging from hate speech to abusive language—in real time. Fine-tuned on a diverse multilingual dataset, it promotes inclusivity by catering to less-resourced languages and providing a toxicity score for each comment to facilitate moderation. This innovative solution offers a robust and scalable method for cultivating healthier and more respectful online communities worldwide.

Citation**:** ***A. Akshaya, K. Sindhuja, N. Rohan and Y. Sahas 2025.*** *"*Multilingual Toxic Comments Classification using Bert*". International Journal of Development Research*, 15, (02), 67737-67742.

INTRODUCTION

The rise of digital communication has transformed how we share ideas, connect with others, and engage in discourse. While this evolution has opened up new avenues for expression, it has also given rise to a troubling increase in toxic behaviours, including hate speech, bullying, and harassment. Such toxic comments can create hostile environments, deterring individuals from participating in discussions and ultimately compromising the integrity of online communities. Consequently, many platforms have struggled with effective content moderation, often resorting to restricting or disabling user comments, which hinders open and constructive dialogue.Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to evidently facilitate conversations, leading many communities to limit or completely shut down user comments.The rise of online platforms has fundamentally changed how we communicate across the globe, allowing individuals from varied linguistic and cultural backgrounds to connect and share ideas. Unfortunately, this greater accessibility has also led to an increase in toxic comments—ranging from hate speech to harassment and abusive language—which can damage individuals and disrupt healthy dialogue. While there are systems in place to tackle these issues, they often focus on English and overlook the significant challenges posed by toxic behaviour in a multilingual context. As a result, many people worldwide lack adequate safeguards against harmful interactions. To effectively combat multilingual toxicity, we need solutions that can understand and analyze text across different languages, each with its own unique grammar, syntax, and cultural context. Traditional machine learning approaches frequently falter in this area due to a lack of annotated datasets for less-resourced languages, highlighting an urgent need for innovative methods that can reconcile linguistic diversity with effective moderation. This project utilizes Multilingual BERT (mBERT), a pre-trained transformer model, to classify toxic comments in various languages. By fine-tuning mBERT on a multilingual dataset, the model is able to recognize toxic patterns, providing support for both high-resource and low-resource languages alike. Additionally, the system generates toxicity scores to facilitate nuanced moderation, allowing online platforms to implement specific interventions based on the severity of the content. Leveraging mBERT promotes inclusivity and scalability, making it an instrumental resource for addressing online toxicity in our increasingly interconnected society. This project not only pushes the boundaries of natural language processing but also aids in fostering safer and more respectful digital communities around the world.Moderating such content poses significant challenges, as manual moderation is time-consuming and inadequate due to the sheer volume of comments.

LITERATURE SURVEY

BERT's pre-trained contextualized embeddings enable accurate toxicity classification, achieving high F1scores. Its bidirectional training considers both the left and right context, improving toxicity detection by capturing nuanced relationships between words. With a deep learning architecture that learns complex patterns in text data, BERT adapts well to various toxicity classification tasks. Pre-trained on vast amounts of text data, BERT reduces false positives and false negatives, ensuring robust performance. This effectiveness fosters safer online environments by identifying toxic comments with precision. Overall, BERT's advanced capabilities make it an ideal solution for toxicity classification [1]. The classification of toxic comments has been an important focus in the field of natural language processing (NLP), primarily due to its significance in fostering healthy online interactions. Initially, research in this area leveraged rule-based systems and conventional machine learning algorithms, notably Support Vector Machines (SVM) and Naive Bayes. These early methods utilized handcrafted features, including bag-of-words and TF-IDF. While they established a foundational understanding, they often struggled to effectively capture the complex and contextual characteristics of toxic language, particularly when faced with challenges such as sarcasm, slang, or the dynamic nature of toxicity patterns [2]. The emergence of deep learning has marked a pivotal shift in the field, particularly with the utilization of models like Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for text classification. These models have demonstrated notable improvements in grasping semantic relationships within text, especially when leveraged alongside pre-trained word embeddings such as Glove and Word2Vec. Nonetheless, it is important to highlight that these methodologies have largely been constrained to specific languages, grappling with challenges in multilingual environments due to their dependence on monolingual embeddings and datasets [3]. The classification of toxic comments has been extensively explored within the field of natural language processing (NLP), driven by the need to foster healthy online environments. Initial methodologies predominantly employed rule-based systems along with traditional machine learning techniques such as Support Vector Machines (SVM) and Naive Bayes. These systems used curate features like bag-of-words and TF-IDF. Although these strategies laid the groundwork for toxic comment classification, they often fell short in capturing the complexity and context of toxic expressions, particularly with nuances like sarcasm, slang, and the ever-changing nature of toxic behaviour. The emergence of deep learning has significantly transformed this landscape, with models such as Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) gaining prominence for text classification initiatives. When integrated with pre-trained word embeddings like Glove or Word2Vec, these advanced models have enhanced the comprehension of semantic relationships within text. Nonetheless, a limiting factor remains as these techniques tended to be language-specific, facing challenges in multilingual contexts due to their dependence on monolingual embeddings and datasets[4][5].

The introduction of transformer-based models, particularly Multilingual BERT (mBERT), revolutionized multilingual toxic comments classification. mBERT leverages a shared vocabulary and a single model architecture to learn representations for over 100 languages, enabling cross-lingual transfer learning. Researchers have demonstrated that fine-tuning mBERT on multilingual datasets significantly improves accuracy and scalability. Moreover, mBERT's contextual embeddings capture linguistic and cultural nuances, addressing challenges like code-switching and mixed-language text. This advancement underscores the potential of transformer-based models in developing inclusive and effective solutions for multilingual toxicity detection [6]. Multilingual toxic comment classification has become an essential focus of research, underscoring the necessity to tackle toxicity within various linguistic frameworks. Initial investigations in this field employed conventional machine learning techniques, such as Support Vector Machines (SVMs) and Naive Bayes, by training these models on multilingual datasets. However, these approaches depended on manually crafted features and demonstrated limited ability to generalize across languages, particularly those with scarce annotated resources. This shortcoming emphasized the demand for more effective methodologies capable of addressing linguistic diversity and subtleties.The emergence of deep learning has facilitated the utilization of embeddings such as fast Text and multilingual word2vec for text representation across multiple languages. When integrated with deep neural networks like LSTMs and CNNs, these embeddings significantly enhanced the efficacy of multilingual toxicity detection. Nevertheless, despite these improvements, the models frequently encountered difficulties in cross-lingual comprehension, as they were predominantly reliant on language-specific training datasets. Additionally, numerous languages suffered from a lack of adequate annotated datasets, complicating the development of precise classifiers for languages with fewer resources [7][8]. The classification of toxic comments in multiple languages has become a critical concern due to the widespread influence of social media and online platforms. Initial attempts in this domain employed conventional models such as Naive Bayes, Decision Trees, and Support Vector Machines (SVMs), along with manually crafted features like n-grams and TF-IDF. Although these methods offered fundamental capabilities for identifying toxic comments, they faced challenges with multilingual datasets, primarily due to their limited semantic comprehension and dependence on language-specific pre-processing techniques. Furthermore, these approaches were inadequate in addressing intricate linguistic elements, including context, sarcasm, and implicit forms of toxicity [9]. The advent of deep learning has led to the development of techniques utilizing word embeddings such as Glove, Word2Vec, and fast Text to improve the representation of multilingual text. These embeddings facilitate models in grasping the semantic connections between words in various languages. When integrated with advanced architectures like Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), these techniques have markedly enhanced performance. Nonetheless, they continue to encounter difficulties, especially when dealing with languages that have scarce training data or those that necessitate a nuanced comprehension of cultural and linguistic distinctions.Transformers have represented a significant advancement in the classification of multilingual toxicity. Models such as Multilingual BERT (mBERT) and XLM-RoBERTa provide effective solutions by employing a shared vocabulary and being pre-trained on extensive multilingual datasets. These models utilize self-attention mechanisms to comprehend context, rendering them particularly adept at identifying toxicity across a variety of languages. Research indicates that fine-tuning mBERT on multilingual datasets yields superior performance compared to conventional methods, even for languages with limitedresources [10] [11].

EXISTING SYSTEM

The development of multilingual systems for toxic comment classification has significantly evolved to address the challenges posed by toxicity in diverse linguistic contexts. Early models in this area were based on traditional machine learning approaches, such as Support Vector Machines (SVMs), Naive Bayes, and Logistic Regression. These models were typically trained using manually engineered features like n-grams, TF-IDF, or syntactic structures. Although these systems were able to perform basic toxic comment detection, they were limited in their ability to handle the complexities of language, such as idiomatic expressions, polysemy, and contextual meaning. Furthermore, they struggled with scalability as they often required separate models for each language, leading to performance degradation when applied to a wide variety of languages or multilingual datasets. Additionally, many early systems lacked the ability to account for linguistic diversity and cultural context, which is essential when detecting toxicity in user-generated content.

Early Approaches: Traditional Machine Learning Models Early systems for multilingual toxic comment classification primarily relied on traditional machine learning models, such as SVMs, Naive Bayes, and Logistic Regression. These models used manually crafted features like n-grams and TF-IDF, but they struggled with scalability, accuracy, and the ability to handle complex language structures across different languages.

Deep Learning Advancements: Word Embeddings and RNNs The adoption of deep learning techniques, particularly word embeddings like Word2Vec and fast Text, helped improve the semantic understanding of words in different languages. Coupled with

RNNs, LSTMs, and CNNs, these models began to capture better language representations and context, though they still struggled with handling mixed-language texts and nuanced contexts.

Breakthrough with Transformer-Based Models: Transformer-based models like Multilingual BERT (mBERT) and XLM-RoBERTa revolutionized multilingual toxic comment classification. By leveraging a bidirectional attention mechanism, these models could capture contextual relationships in text and generalize well across multiple languages, significantly improving performance over previous approaches.

Cross-Lingual Transfer Learning and LowResourceLanguages A major advantage of transformer models is their ability to handle low-resource languages through cross-lingual transfer learning. These models use pre-trained knowledge from high-resource languages and fine-tune on multilingual datasets, improving accuracy even in languages with limited annotated data.

Challenges with Data Imbalance and Mixed-LanguageContent Despite improvements, multilingual models still face challenges related to data imbalance and handling mixed-language or code-switched content. Techniques like data augmentation and cross-lingual attention mechanisms are being explored to address these issues, but they remain ongoing challenges.

Ethical Concerns and Fairness in Multilingual Systems: Ethical considerations and fairness are critical in multilingual toxicity detection. Transformer models can introduce biases due to the underrepresentation of certain languages or cultures, which could lead to unequal performance. Researchers are working on methods to ensure fairness, reduce bias, and improve inclusivity in multilingual systems.

Future Directions and Ongoing Research: Ongoing research continues to focus on improving multilingual toxic comment classification models by enhancing fairness, accuracy, and the handling of mixed-language texts. Methods like adversarial training, bias detection, and better data representation are actively being explored to ensure that multilingual models can effectively detect toxicity while remaining unbiased and culturally sensitive.

PROPOSED SYSTEM

Toxicity Scoring: The proposed system will include a toxicity scoring mechanism that assigns a score to each comment based on its level of toxicity. This score will help prioritize which comments need immediate attention. The score will be derived from the output of the model's classification head, which evaluates the likelihood that a comment is toxic. The higher the toxicity score, the more likely it is that the comment contains harmful content. This quantitative assessment allows for more nuanced moderation and enables the system to flag high-priority content efficiently.

Accuracy: Accuracy will be a key performance metric for evaluating the system's effectiveness. Given the multilingual nature of the task, the model's accuracy will be measured across different languages and comment types. The accuracy will be determined by comparing the predicted toxicity labels with the ground truth labels in the test set. Additionally, metrics such as precision, recall, and F1-score will be used to assess the model's ability to correctly identify both toxic and non-toxic comments. Ensuring high accuracy across various languages and contexts is critical for the success of the system in real-world applications.

METHODOLOGY

Data Collection and Pre-processing:

- ***Dataset:*** Use a large labeled dataset of toxicand non-toxic comments.
- ***Cleaning:*** Remove special characters, correct spelling, and tokenize text.
- ***Augmentation:*** Generate additional data using techniques like synonym replacement and back translation.
- ***Balancing:*** Address class imbalance with under-sampling or over-sampling
- ***Model Selection and Fine-tuningBERT:*** Use a pre-trained BERT model for fine-tuning on the dataset.
- ***Hyper parameter Tuning:*** Optimize learning rate, batch size, and epochs.
- ***Cross-Validation:*** Apply k-fold cross-validation to ensure model robustness.
- ***Toxicity Scoring SystemScore Generation:*** The model outputs a continuous toxicity score (0 to 1).
- ***Threshold:*** Apply dynamic or context-aware thresholds to classify toxicity levels.

Model Evaluation:

- **Metrics**: Measure accuracy, precision, recall, F1-score, and AUC-ROC.
- **Error Analysis**: Review misclassifications to improve the model.

Deployment and Real-Time Prediction:

- **Backend**: Deploy using Flask for real-time comment classification.
- **Optimization**: Compress the model for faster inference and use caching for frequent comments

Continuous Improvement

- ***Monitoring:*** Continuously monitor performance and retrain with new data.
- ***User Feedback:*** Incorporate user feedback to enhance the model's accuracy.

EXPERIMENTAL RESULTS

Distribution of Toxic Labels (Multilingual): The bar chart titled "Distribution of Toxic Labels" highlights the prevalence of different toxic comment categories across various languages. The dataset includes six primary types of toxic comments: toxic, severe_toxic, obscene, threat, insult, and identity hate. By aggregating the frequency of each label across multilingual data, we gain insights into the distribution of these toxic behaviors in different linguistic contexts.

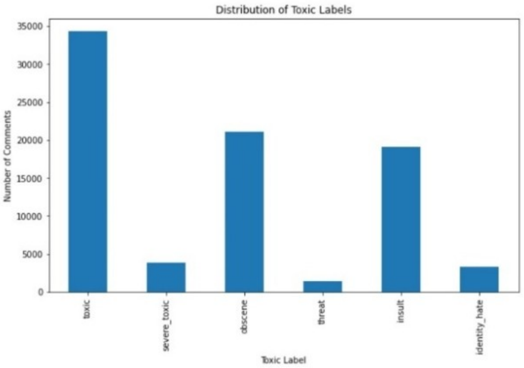
- ***Toxic Comments:*** 35,000 comments in total, the most prevalent type across languages.
- ***Obscene Comments:*** 20,000 comments, showing a notable presence of vulgar or explicit content.
- ***Insults:*** 18,000 comments, reflecting a common form of harmful language.
- ***Severe Toxic Comments:*** 5,000 comments, indicating extreme toxicity with significant impact.
- ***Identity Hate:*** 2,500 comments, representing harmful content targeting specific groups.
- ***Threats:*** 1,000 comments, the least frequent but still critical in severity.

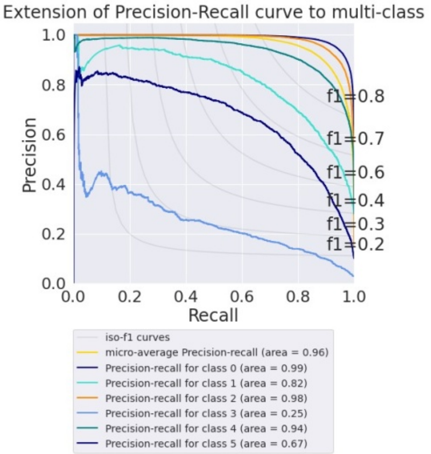
These statistics reveal the most common forms of toxicity and suggest that the prevalence of toxic comments varies by language. This can guide tailored moderation efforts and improve the multilingual toxicity classification model.

Next Steps

Develop Classification Models: Use this data to train more accurate models for detecting and managing toxic comments.

Content Moderation Strategies: Tailor moderation efforts to address the most prevalent forms of toxicity.





Distribution of Toxicity Types (Multilingual): The bar chart "Distribution of Toxicity Types" provides an overview of the prevalence of various toxic comments across multiple languages. This distribution reveals the following key insights:

- ***Toxic Comments:*** 34,738 comments, forming the bulk of the dataset in multiple languages.
- ***Severe Toxic:*** 1,874 comments, less frequent but highly damaging.
- ***Obscene:*** 20,827 comments, indicating a high frequency of explicit content across languages.
- ***Threat:*** 92 comments, the least frequent but crucial in severity.
- ***Insult:*** 18,590 comments, a substantial portion of harmful comments in multiple languages.
- ***Identity Hate:*** 1,315 comments, showing a notable presence of targeted harassment in different languages.

These insights suggest that while some forms of toxicity, like toxic, obscene, and insulting comments, dominate across multiple languages, less frequent but significant toxicity types, such as threats and identity hate, still require attention.

Distribution of Toxic Labels

Bar Chart: The bar chart titled "Toxicity Distribution of Severe Comments" displays the count of different types of toxic comments categorized into six types:
Toxic:1800 comments
Severe_toxic: 1800 comments
Obscene: 1800 comments
Threat: 100 comments
Insult: 1800 comments
Identity_hate: 500 comments.

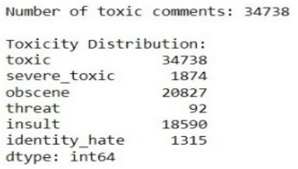
Key Insights:The toxic, severe_toxic, obscene, and insult categories have similar high frequencies, each with around 1800 comments. This highlights that these forms of toxicity are quite prevalent in the dataset.Identity_hate is present but less common, with around 500 comments.Threats are the least frequent, with approximately 100 comments.

Impact and Application: This visualization aids in understanding the distribution of various toxic labels. It is crucial for developing strategies to enhance classification models and content moderation efforts. By identifying the most common forms of harmful comments, we can target interventions to mitigate toxic behavior in online environments effectively.

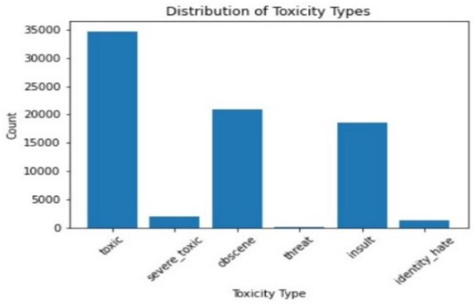
Correlation Between Toxicity Types (Multilingual): A heatmap illustrating the correlation between different toxicity types helps us understand how these toxic behaviors interrelate across multilingual datasets. The key correlations observed are:

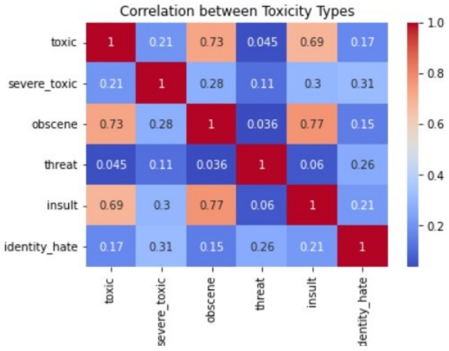
- **Toxic and Obscene (0.73):** These two forms of toxicity are strongly linked across different languages. Comments that are labeled as toxic often contain obscene content, regardless of the language.
- **Toxic and Insult (0.69):** A strong connection is found between general toxicity and insults, suggesting that personal attacks often co-occur with toxic comments in multiple languages.
- **Obscene and Insult (0.77):** The highest correlation is found between obscenity and insults, revealing that these two forms of toxic behavior frequently appear together in harmful comments.
- **Severe_toxic and Insult (0.3):** A moderate correlation exists between severe toxicity and insults, indicating that while the two often appear together, they are not as strongly linked as other types.
- **Threat and Identity Hate (0.26):** A weak but notable correlation between threats and identity hate, indicating that while both types of toxicity are less common, they can occasionally overlap.

These correlations inform more effective multilingual content moderation strategies by highlighting the co-occurrence of toxic behaviors and helping refine the classification system to better handle these complex interactions.


```
Number of toxic comments: 34738

Toxicity Distribution:
toxic            34738
severe_toxic      1874
obscene          20827
threat              92
insult           18590
identity_hate     1315
dtype: int64
```





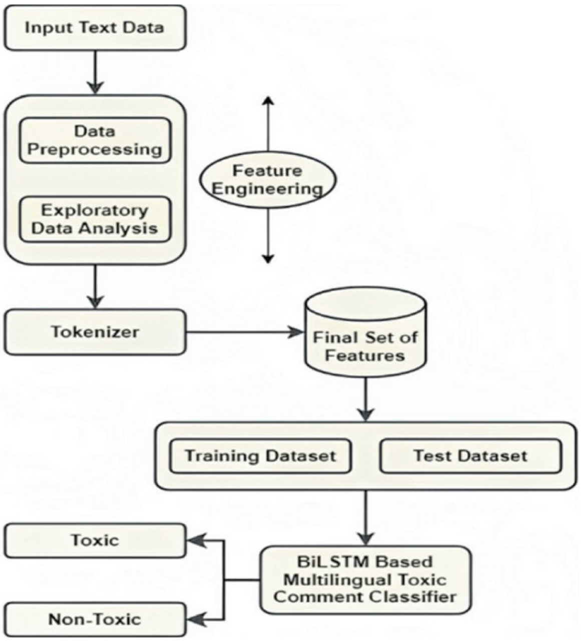
Occurrences of Various Toxicity Types in Severe Toxic Comments (Multilingual)

A bar chart titled "Occurrences of Various Toxicity Types in Severe Toxic Comments" reveals the distribution of different toxicities within the severe toxic category, with a multilingual focus. The insights are as follows:

- **Obscene:** 1,750 comments, showing a dominant presence of explicit language in severe toxicity cases.
- **Insult:** 1,750 comments, equally frequent, indicating that severe toxicity often involves personal attacks.
- **Identity Hate:** 500 comments, still a significant concern but less common compared to other forms.
- **Threats:** Minimal presence, with near-zero occurrences, suggesting that severe toxicity rarely involves threats.

This analysis highlights that obscene language and insults dominate severe toxic comments across multiple languages, emphasizing the need for focused moderation strategies. Identifying these patterns allows for better detection and mitigation of severe toxic content in multilingual environments.

$$\frac{Number\ of\ correctly\ classified\ comments}{Total\ number\ of\ comments} = 0.8999307931367684$$



CONCLUSION

The Multilingual Toxic Comments Classification system aims to automatically detect and categorize harmful comments in multiple languages, enabling better content moderation across diverse online platforms. This system classifies toxic comments into several categories, including toxic, severe toxic, obscene, threat, insult, andidentity hate. By employing advanced Natural Language Processing (NLP) techniques, the model can handle a variety of languages, making it effective in moderating content in global online communities.In this approach, the BERT (Bidirectional Encoder Representations from Transformers) model is used for understanding the context of words in a sentence by capturing both the preceding and succeeding words, making it suitable for multilingual classification. The model is pre-trained on large datasets and then fine-tuned with a multilingual corpus that helps it grasp the nuances of different languages while detecting harmful behaviors in text.

The system performs several key tasks, such astext preprocessing, feature extraction, and classification. First, the text data is cleaned, and relevant features are extracted to help the model understand the underlying toxicity. Then, the pre-trained multilingual BERT model is fine-tuned on the labeled toxic comment dataset to accurately classify comments into the appropriate categories.By leveraging multi-class classification for some labels and multi-label classification for others, this system ensures that comments can be classified into multiple categories if needed. For instance, a comment can be both toxic and insulting, and the system is capable of handling these overlapping labels efficiently. This ability to detect multiple toxic behaviors simultaneously is crucial for effective content moderation.The model's performance is evaluated using accuracy, precision, recall, and F1-score metrics, ensuring that it not only identifies toxic comments but also minimizes false positives and false negatives. The insights derived from the classification results, such as the frequency of each toxicity type across different languages, inform content moderation strategies and help address the most prevalent forms of toxic behavior online.In summary, the multilingual toxic comments classification system is a powerful tool for automating the detection of harmful content in various languages, improving user experience, and promoting a safer online environment across global platforms

REFERENCES

Akhil Kumar, K.G.S.S.S.V., Kanisha, B.: Analysis of multiple toxicities using ML algorithms to detect toxic comments. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), India, IEEE Explore, July (2022). https://doi.org/10.1109/ ICACITE53722.2022.9823822

Gladwin, E. V. Renjiro, B. Valerian, I. S. Edbert and D. Suhartono, "Toxic Comment Identification and Classification using BERT and SVM," 2022 8th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICST56971.2022.10136295. keywords: {Support vector machines;Measurement;Machine learning algorithms;Social networking (online);Bit error rate;Transformers;Natural language processing;Machine Learning;Toxic Comments;Support Vector Machine;Natural Language Processing;Transformer Model},

Jain, S., Kaushik, G., Prabhu, P., Godbole, A.: Detox: NLP based classification and euphemistic text substitution for toxic comments. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), India, pp. 54–61 (2021)

Rupapara, V., Rustam, F., Shahzad, H.F., Mehmood, A., Ashraf, I., Choi, G.S.: Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. IEEE Access 9, 78621–78634 (2021)

S. K. Putri, A. Amalia and T. F. Abidin, "Sentiment Analysis Multi-Label of Toxic Comments using BERT-BiLSTM Methods," 2024 International Conference on Electrical Engineering and Informatics (ICELTICs), Banda Aceh, Indonesia, 2024, pp. 120-124, doi: 10.1109/ICELTICs62730.2024.10776338. keywords: {Electrical engineering; Sentiment analysis; Social networking (online); Semantics; Blogs; Market research;Vectors; Encoding; Internet; Informatics; BiLSTM; BERT; Multi-label; Sentiment analysis},

Stavropoulos ACrone DGrossmann I 2024. Shadows of wisdom: Classifying meta-cognitive and morally grounded narrative content via large language modelsBehavior Research Methods10.3758/s13428-024-02441-056:7(7632-7646) Online publication date: 29-May-2024https://doi.org/10.3758/s13428-024-02441-0

Sumanth, P., Samiuddin, S., Jamal, K., Domakonda, S., Shivani, P.: Toxic speech classification using machine learning algorithms. In: International Conference on Electronic Systems and Intelligent Computing (ICESIC), India, June 2022, IEEE Explore. https://doi.org/10.1109/ICESIC53714.2022.9783475

Tarun, V. G. R. Sivasakthivel, G. Ramar, M. Rajagopal and G. Sivaraman, "Exploring BERT and Bi-LSTM for Toxic Comment Classification: A Comparative Analysis," 2024 Second International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2024, pp. 1-6, doi: 10.1109/

ICDSIS61070.2024.10594466. keywords: {Accuracy; Toxicology; Computational modeling; Oral communication; Transformers; Natural language processing; Real-time systems; toxicity; BERT; bi-LSTM; natural language processing},

Vichare, M., Thorat, S., Uberoi, S., Khedekar, S., Jaikar, S.: Toxic comment analysis for online learning. In: 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), India, IEEE Explore, October 2021. https://doi.org/10.1109/ACCESS51619.2021.9563344

Z. Zhai, "Rating the Severity of Toxic Comments Using BERT-Based Deep Learning Method," 2022 IEEE 5thInternational Conference on Electronics Technology (ICET), Chengdu, China, 2022, pp. 1283-1288, doi: 10.1109/ICET55676.2022.9825384. keywords: {Deep learning; Silver; Gold; Social networking (online); Manuals; Multilayer perceptrons; Internet; Deep Learning;natural language processing; BERT; toxic comment; Kaggle},

Zhang, W., Wu, Y.: Semantic sentiment analysis based on a combination of CNN and LSTM model. In: International Conference on Machine Learning and Knowledge Engineering (MLKE), China IEEE Explore, April 2022. https://doi.org/10.1109/MLKE55170.2022.00041

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2021. A Comparative Study of Using Pre-trained Language Models for Toxic Comment Classification. In Companion Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 500–507. https://doi.org/10.1145/3442442.3452313

****