Enhancing Transparency and Interpretability in Toxic Comment Classification: A Study on the Integration of Explainable Artificial Intelligence (XAI) Techniques

Mr. Jadyn Dias[1]
Student,
Department of MSc. IT,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
jd.jadyndias@gmail.com

Dr. Pallavi Devendra Tawde [2]
Assistant professor,
Department of BSc. IT and CS,
Nagindas Khandwala College,
Mumbai, Maharashtra, India
pallavi.tawde09@gmail.com

Abstract

More than ever, robust, and interpretable toxic comment recognition methods are required to manage the growing frequency of toxic comments on online platforms. The research tries to incorporate techniques in Explainable Artificial Intelligence (XAI) to improve the transparency and comprehensibility of toxic comment classification. Using a comprehensive dataset, we designed a model architecture which includes the latest practices in XAI. Through rigorous experimentation, our study proves the usefulness of such methods as tools that not only increase classification accuracy but also illuminate model decision-making processes. One view is that by adding LIME and Eli5 to toxic comment classification, model performance improves both in terms of accuracy and interpretation for decisions. Our results provide valuable insights into the model's strengths and areas for refinement, contributing to the transparency and interpretability of toxic comment classification. This research contributes to the evolving landscape of interpretable machine learning, offering a pathway to more accountable and trustworthy toxic comment moderation systems.

Keywords: Explainable artificial intelligence, Model interpretability, toxic comment classification, LIME, Eli5

1. Introduction:

Online platforms provide a space for users to share opinions and engage in discussions, thus becoming essential to communication. However, with the drastic increase in dependability on such online platforms, there has been a significant rise in toxic comments which poses a challenge to maintaining a healthy online environment. Automatic classification of toxic comments, such as hate speech, threats, and insults, can help in keeping discussions fruitful [1]. Toxic comments can cause serious consequences and implications including but not limited to cyberbullying, mental health issues, and a decline in user engagement.

To address this problem, the development of robust toxic comment classification models has gained significant importance. Such models aim to identify and filter out toxic comments automatically on their own, therefore fostering a safer and more promising community on these online platforms. The overall effectiveness and reliability of these models is crucial for evaluating and moderating content posted to online platforms and protecting users and their well-being.

While the deployment of AI models for toxic comment classification is essential, it introduces challenges of transparency and interpretability. Users of the platform, content moderators, and all relevant stakeholders need to comprehend why a specific comment is classified as toxic. The "black box" nature of many AI models interferes with this understanding, raising concerns about biased decisions, lack of accountability, and potential unintended consequences. there is a growing concern regarding how machine learning algorithms are learning from the data and making their decisions [2].

The reasons why the research is being performed is due to the need to cover this interpretability gap. The recent development of explainable AI addresses this gap due to the ability to determine why a certain prediction was achieved. Thus, regarding the toxic comment classification, explainability can be a central measure to ensure the reliability of users and their understanding of the development premise as well as their ability to contest the developed solution.

Need for Explainability:

Explainable AI (or XAI) is a new facet of artificial intelligence where we can seek answers to a pressing question of "why?" which is not possible traditionally [3]. The need for explainability in toxic comment classification rises from the underlying complexity of toxic comments, which display a variety of language nuances and context. The complex nature of toxic language makes it challenging to understand the decision-making process of machine learning models without transparent and interpretable explanations.

Explainability plays a major role in promoting accountability and fairness in content moderation. User trust and engagement are closely involved to the transparency of online platforms. When users understand why a comment is labelled as toxic or unhealthy and why a comment is labelled as clean, they are more likely to react and engage positively, contributing to a healthier and safer online environment.

Clear insights into feature contributions are provided by machine learning models that are easy to grasp and interpret, like logistic regression, which improves user comprehension. Explainability plays a crucial role in reducing model biases. Transparent models enable the identification and refinement of biases, enhancing the overall fairness of content moderation. Many hate speech detection models exhibit bias towards specific slurs that are more frequently used against certain groups, leading to the inaccurate classification of hate speech [4].

It is probable that models that strike a balance between transparency and accuracy will be given priority in the field of content moderation as it develops. Incorporating explainability is not optional; rather, it is essential to building a more secure, accountable, and welcoming online community. Content moderation systems have the potential to foster a more responsible and user-focused digital environment by advocating for the need of explainability.

Objective:

The primary goal of our research is to investigate and implement Explainable AI techniques in the context of toxic comment classification. Specifically, we aim to enhance the interpretability of AI models deployed for content moderation on online platforms. By utilizing methods such as Local Interpretable Model-agnostic Explanations (LIME) and Explain Like I'm 5 (Eli5), we aim to propose human-interpretable explanations for the decisions made by toxic comment classifiers. Through this research paper, we aim to contribute to the development of more transparent and accountable AI models in the domain of online platforms.

2. Review of Literature:

Toxic comment classification has been a focal point in natural language processing (NLP) research owing to the rising concerns about online toxicity. Traditional methods often relied on feature engineering and handcrafted rules to identify toxic language.

Qureshi et al. explored the use of machine learning and Explainable AI (XAI) techniques to address the issue of hate speech and offensive language on social media platforms, with a focus on Twitter [4].

Mosca et. al. employed an interpretable deep learning method to categorize hate speech, utilizing the Davidson dataset. The research utilized SHAP values for phrase explanation and incorporated contextual explanations for hate speech classification [6].

The application of XAI in NLP tasks has gained traction, especially in sentiment analysis and text classification. Kumar et al. utilized random forest and extreme gradient boosting (XGBoost) algorithms to identify sarcasm in dialogues. They employed LIME and SHapley Additive exPlanations (SHAP) techniques for result interpretation, allowing users to comprehend the model's decision-making process in sarcasm detection within dialogues more easily [7].

Most previous studies on toxicity detection have prioritized improving model performance, neglecting the importance of explainability. However, Mathew and colleagues introduced a benchmark annotated dataset with explanations and developed an interpretable model for detecting hate speech [8].

Lately, there has been growing attention towards the interpretability of artificial intelligence methods such as machine learning and deep learning, aimed at comprehending the rationale behind categorizing text as hate speech or for other purposes in social media and medical contexts.

A novel explanation method based on LIME for the explanation of predictions made by a classifier was proposed [9], and the best practices for the usage of these interpretable machine learning models were also discussed [10–14].

3. Materials and Methods:

Our strategy involves enhancing the interpretability of toxic comment classification models using Explainable Artificial Intelligence (XAI) techniques. We utilize Local Interpretable Model-agnostic Explanations (LIME) and Explain like I'm 5 (Eli5) to reveal the decision-making processes of our model. By integrating these XAI methods with a state-of-the-art toxic comment classification model, we aim to provide transparent insights into prediction rationale, increasing user trust.

3.1. Kaggle Dataset:

The Kaggle Jigsaw Toxic Comment Classification Challenge dataset is a benchmark for identifying toxic comments in online discussions. It includes diverse comments labeled for various toxicities (toxic, severe toxic, obscene, threat, insult, identity hate). With a multi-label classification task, the dataset mimics real-world scenarios.

3.2. Extracting the Dataset:

We obtained datasets in CSV format, organizing data efficiently. The first row typically serves as the header, containing the column or attribute names. Utilizing the Pandas library in Python, we managed and analyzed the tabular data.

3.3. Data Preprocessing and Cleaning:

Data preprocessing is a crucial phase in our research, laying the groundwork for effective model performance. The Kaggle Toxic Comment Classification dataset, our raw data source, may contain noise, null values, and complexities. Our meticulous preprocessing refines the data, removing noise and retaining meaningful information. The following steps detail our dataset preprocessing:

1. Segregating Clean and Toxic Comments: We separated the dataset into two subsets, clean and toxic comments, based on specific toxicity labels.
2. Crafting Test Set: We created a dedicated test set by independently extracting 20% of clean and toxic comments. These subsets were combined, shuffled, and irrelevant columns removed for enhanced utility.
3. Creating Train Set: The remaining clean and toxic comments were combined for the training set. The set was shuffled, and irrelevant columns were discarded for improved efficiency.

Data cleaning is vital for model training, enhancing data quality by meticulously removing incorrect or inconsistent information. Our approach involves multiple steps, addressing structural errors, outliers, and missing data. Key data cleaning steps include:

1. Text Cleaning Function: We designed a specialized function covering tasks such as converting text to lowercase, removing square brackets, eliminating links, stripping punctuation, and filtering out words with numbers.
2. Application to Datasets: The cleaning function was systematically applied to both training and test datasets comment_text columns, ensuring uniformity and preparing textual data for analysis and model training.

3.4. Feature Engineering:

Text tokenization and vectorization, fundamental in natural language processing (NLP), transform raw textual data into a machine-learning-friendly format. Tokenization breaks sentences into smaller units (words or n-grams), while vectorization represents these units as numerical vectors for machine learning algorithms. To address class imbalance in toxic comment classification, Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic samples, ensuring a balanced dataset. Term Frequency-Inverse Document Frequency (TF-IDF) vectorization converts comment_text data into numerical vectors, capturing word or n-gram importance. Logistic Regression, chosen for its simplicity and effectiveness in binary classification, serves as the predictive model for distinguishing toxic and non-toxic comments.

3.5. *Classification and Cross-Validation:*

A logistic regression model is trained with specific hyperparameters, and its performance is assessed through cross-validation using F1-weighted scoring. After cross-validation, the model is trained on the entire resampled dataset for deployment on new instances.

3.6. Explainability Methods:

1. Local Interpretable Model-agnostic Explanations (LIME): Complex models like deep neural networks lack interpretability, hindering a clear understanding of predictions. LIME addresses this by providing local, interpretable explanations for individual instances, unravelling complexities in a way humans can grasp. When identifying toxic comments, LIME elucidates pivotal words or features, offering actionable insights through visualizations. This interpretability is crucial for trust and allows stakeholders to intervene when necessary.

2. Eli5 Integration: Integrated for debugging and interpreting machine learning models, Eli5 provides global and feature-level explanations. Global explanations offer an overview of the model's behavior, while feature-level explanations highlight specific words or tokens in toxicity classification. Eli5 bridges the gap between complex algorithms and transparency, fostering trust and understanding. It acts as a friendly guide in the machine learning landscape, ensuring accurate and explainable model decisions.

4. Results and Analysis:

4.1. Model Evaluation:

To assess our toxic comment classification model, we conducted a thorough evaluation using a test set, employing a custom function to generate key metrics and reports. The classification report includes precision, recall, and F1-score for clean and toxic classes, providing detailed insights into the model's performance. The accuracy score offers a global perspective on correctness. These metrics contribute to understanding the model's strengths and limitations, enhancing transparency and interpretability in our research.

4.2. Result Analysis:

The results of the toxic comment classification model are presented below:

The overall accuracy of the model is 89.7%, showcasing its ability to make correct predictions across both classes. The weighted average F1-score is 91%, demonstrating a balanced trade-off between precision and recall. The macro average F1-score is 77%, indicating a moderate level of performance across classes.

These results indicate that the model is effective in distinguishing between clean and toxic comments, with notable precision in identifying clean comments and reasonable performance in capturing toxic instances. The findings provide valuable insights into the model's strengths and areas for potential refinement in future iterations.

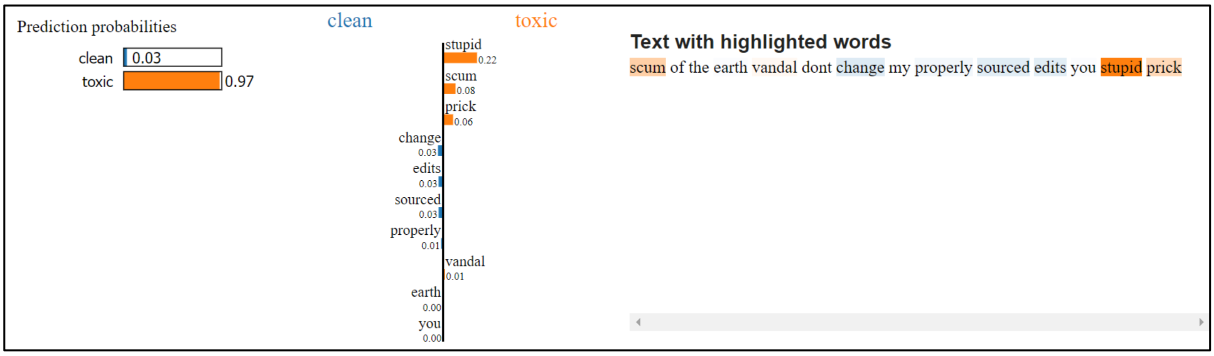
	precision	recall	f1-score	Support
clean	0.98	0.90	0.94	28669
toxic	0.48	0.83	0.61	3059
Macro avg	0.73	0.87	0.77	31728
Weighted avg	0.93	0.90	0.91	31728

Table 1: Result Analysis

4.3. Explainability with LIME:

LIME analysis offers crucial insights into the interpretability of our toxic comment classification model. For a specific comment, the model predicts a high toxicity probability of 0.97, confidently classifying it as toxic. Highlighted words like "scum," "vandal," "stupid," and "prick" significantly contribute to this prediction, each with associated probabilities.

This analysis underscores the model's focus on specific words indicative of toxicity, providing a transparent view of influential features. The model correctly identifies and assigns high importance to negative terms, enhancing our understanding of its decision-making process. LIME's interpretability aids in building trust and understanding, offering a valuable tool for users and stakeholders to comprehend and validate predictions.



Text with highlighted words

scum of the earth vandal dont change my properly sourced edits you stupid prick

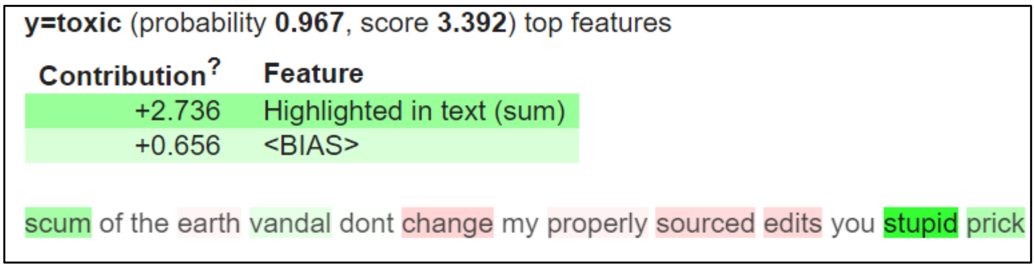
Figure 1: Explainability with LIME

4.4. Explainability with Eli5:

Eli5 analysis offers insightful interpretations of our toxic comment classification model, highlighting key features influencing its decision-making. Specifically, for a toxic comment, the model predicts high toxicity (0.967) with a score of 3.392.

The most impactful feature is the sum of weights assigned to highlighted words, contributing significantly with a score of +2.736. Words like "scum," "vandal," "stupid," and "prick" strongly influence the toxic classification. The bias term (<BIAS>) also plays a notable role with a positive score of +0.656, representing the baseline influence on predictions.

Understanding these top features is crucial for grasping the model's sensitivity to language patterns linked with toxicity. Eli5 demystifies the model's black-box nature, providing transparency and actionable insights into the elements contributing most to toxic classification. This enhanced interpretability is vital for building trust in the model's decisions, particularly in applications prioritizing transparency and accountability, such as toxic comment moderation.



y=toxic (probability **0.967**, score **3.392**) top features

Contribution?	Feature
+2.736	Highlighted in text (sum)
+0.656	<BIAS>

scum of the earth vandal dont change my properly sourced edits you stupid prick

Figure 2: Explainability with Eli5

5. Discussion:

5.1. Challenges:

During the implementation and analysis of the toxic comment classification model, several challenges were encountered.

1. Data Imbalance: We had to address a significant class imbalance between clean and toxic comments which was crucial to prevent bias towards the majority class. We achieved this by implementing Synthetic Minority Over-sampling Technique (SMOTE) to mitigate this challenge.
2. Model Complexity: We implemented a robust machine learning model capable of accurately classifying toxic comments while ensuring interpretability was a delicate balance. Additionally, fine-tuning of the model's hyperparameters and selection of appropriate features were challenges in achieving this balance.

5.2. Limitations:

Despite the successful implementation, there are certain limitations:

1. Language Nuances: The model may struggle with understanding context-specific nuances and sarcasm in comments, potentially leading to misclassifications.
2. Generalization: The model might face challenges in generalizing well to diverse linguistic styles and evolving online language trends.
3. Overfitting: Despite implementing cross-validation techniques, the model may overfit to the training data, limiting its performance on unseen data.

5.3. Ethical Considerations:

Toxic comment classification holds ethical implications related to content moderation and freedom of expression. It is crucial to strike a balance between identifying harmful content and avoiding censorship. Explainability becomes paramount in this context, allowing users to understand why a comment was classified as toxic. Transparency in the classification process can help mitigate biases and ensure fairness. Additionally, it is essential to regularly update the model to adapt to evolving language norms and prevent the amplification of biases over time. Ethical considerations should be an ongoing part of the development and deployment of such models to align them with societal values and expectations.

6. Conclusion:

The implemented model demonstrated commendable performance in distinguishing between clean and toxic comments, achieving an accuracy of 89.7%. The classification report highlighted the ability to identify toxic comments with reasonable precision and recall. LIME-generated explanations offered localized insights, illustrating the model's reasoning for specific toxic comments. This enhanced interpretability aids users in understanding why certain comments are classified as toxic. ELi5 provided valuable insights into the model's decision-making process. It showcased top features contributing to the toxicity prediction, such as highlighted words in the text and the bias term. This study places a significant emphasis on interpretability, leveraging ELi5 and LIME to demystify the complex decision-making of toxic comment classification models. The developed model has implications for online platforms and social media in efficiently identifying and moderating toxic comments, contributing to a safer online environment. The insights provided by ELi5, and LIME have implications for user understanding. Users can now comprehend why a comment is labelled as toxic, fostering transparency and trust.

7. Future Scope

A primary focus should be on extending the application of Explainable Artificial Intelligence (XAI) techniques to address challenges associated with multimodal data, such as toxic content embedded in images or videos. Exploring methods to combine textual and visual information will significantly enhance the interpretability of models in scenarios involving multimedia elements. Moreover, there is a crucial need to delve into bias mitigation strategies, ensuring fair treatment across diverse user groups and addressing potential sources of bias in model predictions. Linguistic diversity poses another challenge, urging the exploration of explainability methods that effectively handle comments in multiple languages and account for cultural nuances. Investigating language-agnostic approaches will contribute to cross-cultural interpretability. Real-time interpretability solutions are essential for online platforms, requiring research into techniques that balance speed and accuracy in providing instantaneous insights.

References

[1] Van Aken, B., Risch, J., Krestel, R., & Löser, A. (2018). Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572.

[2] Zahoor, K., Bawany, N. Z., & Qamar, T. Evaluating text classification with explainable artificial intelligence. Int J Artif Intell ISSN, 2252(8938), 8938.

[3] Mehta, H., & Passi, K. (2022). Social media hate speech detection using explainable artificial intelligence (XAI). Algorithms, 15(8), 291.

[4] Mazhar Qureshi, M. D., Qureshi, M. A., & Rashwan, W. (2023). Toward Inclusive Online Environments: Counterfactual-Inspired XAI for Detecting and Interpreting Hateful and Offensive Tweets.

[5] Mosca, E., Wich, M., & Groh, G. (2021, June). Understanding and interpreting the impact of user context in hate speech detection. In Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media (pp. 91-102).

[6] Mosca, E., Wich, M., & Groh, G. (2021, June). Understanding and interpreting the impact of user context in hate speech detection. In Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media (pp. 91-102).

[7] Kumar, A., Dikshit, S., & Albuquerque, V. H. C. (2021). Explainable artificial intelligence for sarcasm detection in dialogues. Wireless Communications and Mobile Computing, 2021, 1-13.

[8] Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021, May). Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 17, pp. 14867-14875).

[9] Androcec, D. (2020). Machine learning methods for toxic comment classification: a systematic review. Acta Universitatis Sapientiae, Informatica, 12(2), 205-216.

[10] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247-278.

[11] Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences, 12(3), 1353.

[12] Xingyi, G., & Adnan, H. (2024). Potential cyberbullying detection in social media platforms based on a multi-task learning framework. International Journal of Data and Network Science, 8(1), 25-34.

[13] Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). A diagnostic study of explainability techniques for text classification. arXiv preprint arXiv:2009.13295.

[14] Vilone, G., & Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. Machine Learning and Knowledge Extraction, 3(3), 615-661.

[15] Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89-106.

[16] Saif, M. A., Medvedev, A. N., Medvedev, M. A., & Atanasova, T. (2018, December). Classification of online toxic comments using the logistic regression and neural networks models. In AIP conference proceedings (Vol. 2048, No. 1). AIP Publishing.

[17] Ozoh, P. A., Adigun, A. A., & Olayiwola, M. O. (2019). Identification and classification of toxic comments on social media using machine learning techniques. International Journal of Research and Innovation in Applied Science (IJRIAS), 4(11), 142-147.

[18] Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.

[19] Nelatoori, K. B., & Kommanti, H. B. (2023). Multi-task learning for toxic comment classification and rationale extraction. Journal of Intelligent Information Systems, 60(2), 495-519.

[20] Balkir, E.; Nejadgholi, I.; Fraser, K.C.; Kiritchenko, S. Necessity and sufficiency for explaining text classifiers: A case study in hate speech detection. arXiv 2022, arXiv:2205.03302.

[21] Mahajan, A.; Shah, D.; Jafar, G. Explainable AI approach towards toxic comment classification. In Emerging Technologies in Data Mining and Information Security; Springer: Singapore, 2021; pp. 849–858