An Explainable Learning Framework for Detecting Abusive Language

Soumik Datta
Department of CSE, BUET
Dhaka, Bangladesh
0424052089@grad.cse.buet.ac.bd

Abstract

This research endeavors to enhance the interpretability and operational efficiency of text classification models under adversarial conditions, employing a comprehensive methodology. Initially, the textual data undergoes meticulous preprocessing, followed by the deployment of FastText for sophisticated word embedding, optimizing the trade-off between complexity and performance. The core of the study involves the selection of an optimal lightweight model, characterized by its high efficiency and robustness. Subsequent evaluations under adversarial attacks reveal slight performance decrements, underscoring the model's vulnerabilities. To augment transparency and accountability, the Local Interpretable Model-agnostic Explanations (LIME) technique is integrated, providing detailed insights into the model's reasoning processes. Results indicate that combining FastText embeddings with a lightweight model framework not only achieves superior performance but also enhances interpretability. The integration of LIME further solidifies the model's reliability by elucidating decision-making processes. Future research will focus on strengthening model resilience against a wider array of adversarial strategies and expanding these methodologies to additional text analysis domains, aiming to bolster the application of explainable AI in critical decision-making environments. This study contributes significantly to advancing machine learning models that are both effective and transparent.

Keywords

FastText, LIME, Lightweight,

1 Introduction

The proliferation of social media and online forums has facilitated a surge in the expression of diverse opinions and interactions among users from various backgrounds [9]. However, this increase in digital communication has also led to a significant rise in abusive language, which can manifest as hate speech, cyberbullying, and other forms of online harassment. Detecting and mitigating abusive language is crucial not only for maintaining the quality of discourse but also for protecting individuals from harm [7]. Current automated systems employ various machine-learning techniques to detect abusive content effectively. However, these systems often act as black boxes, providing little to no insight into the reasoning behind their decisions.

This lack of transparency is problematic for several reasons [8]. First, it impedes the trust users and regulators have in these systems, especially in scenarios where justification of decisions is required, such as content moderation in legally sensitive contexts. Second, without understanding the model's decision-making process, developers and researchers cannot easily identify or correct biases that the model may have learned. Biased models can lead to unfair treatment of certain groups or individuals, exacerbating the issues they are meant to mitigate [2]

To address these challenges, our research introduces an explainable learning framework specifically designed to detect abusive language in online platforms. This framework integrates robust text preprocessing, optimized text representations via fastText embeddings, and a lightweight machine learning classifier to efficiently and effectively categorize text. Central to our approach is the incorporation of Local Interpretable Model-agnostic Explanations (LIME), which provides insights into the model's predictions, offering a clear explanation of why specific content is flagged as abusive.

Moreover, recognizing the dynamic nature of language and the continuous evolution of online discourse, we extend our model's robustness through adversarial training. This method tests the model's resilience against inputs crafted to circumvent detection, thereby enhancing its ability to handle novel or evolving abusive expressions. Our approach not only contributes to the field by improving detection accuracy and efficiency but also by reducing the model's memory footprint, making it suitable for real-time applications in resource-constrained environments.

In summary, this paper presents a comprehensive approach to abusive language detection that balances performance, explainability, and computational efficiency. By doing so, it addresses critical gaps in current methodologies and lays the groundwork for more transparent, fair, and effective moderation tools in social media platforms. This research contributes to the broader discourse on ethical AI, advocating for systems that uphold the principles of accountability and fairness in automated content moderation.

2 Related Works

In the field of automated hate speech detection, substantial progress has been made through the application of machine learning (ML) techniques. Initial studies often focused on utilizing single-classifier systems to address specific classes of hate speech detection problems [4]. While these approaches laid the foundational framework, they generally lacked the robustness needed to handle the linguistic variability and often encountered issues with imbalanced datasets [4]

As a result, researchers have explored multi-class classification datasets to enhance the robustness of models. However, some researchers [11] struggled to address the imbalanced nature of these datasets, which often contain a disproportionately low number of samples from abusive classes, leading to biased prediction results.

To improve model performance, various deep learning methods have been employed [12, 13]. These methods, being more context-dependent, have the potential to achieve higher performance. However, they also require substantial computational resources and time [12, 13].

Furthermore, in the pursuit of enhancing model performance, some researchers [13, 14] have overlooked the importance of model transparency. This oversight can create a trade-off between performance and transparency, where increasing one can negatively affect the other.

In the realm of adversarial attacks, a study by [15] introduced a novel method aimed at addressing trustworthiness issues in machine learning text classifiers. Applied to models trained on the HateXplain dataset, this approach uncovered vulnerabilities and underscored the need for robust defense mechanisms.

3 Problem Formulation with Example

3.1 Problem Statement

Effective moderation of abusive language on digital platforms is crucial for fostering healthy interactions. However, automatic detection is hindered by several technical challenges that affect the performance and reliability of these systems. Our research aims to address these challenges through an innovative approach that enhances accuracy, efficiency, transparency, and robustness against adversarial attacks.

3.2 Challenges

Our approach is to detecting abusive language online must navigate several complex challenges, each of which influences the design of our solution:

- **Imbalanced Data:** Abusive language data sets are typically imbalanced, with far fewer examples of abuse than non-abusive content. This skew can lead to models that are biased towards the majority class, resulting in higher rates of false negatives.
- **Computational Efficiency:** Traditional deep learning approaches require significant computational resources, which can be a barrier for real-time processing and deployment in resource-constrained environments.
- **Context Dependence:** The meaning of words or phrases can vary significantly based on context, making it difficult for models that do not account for the surrounding content or the conversation's history to accurately interpret messages.
- **Explainability:** There is a growing demand for models to not only predict accurately but also provide understandable explanations for their decisions, particularly in sensitive areas such as content moderation.
- **Adversarial Attacks:** Models are often susceptible to adversarial attacks, where slight, often human imperceptible alterations to input data can lead to incorrect classifications.
- **Efficient Word Embeddings:** Many existing embedding techniques do not capture semantic nuances well and can be inefficient in terms of storage and computation, particularly when scaling to large vocabularies and datasets.

Figure 1 illustrates a streamlined process from input through preprocessing to prediction and explanation, highlighting the role of machine learning and local explainability in text classification systems. The system is designed to predict whether the text is offensive and provide insights into which words or phrases triggered that classification, thus making it more transparent and understandable for users or moderators.

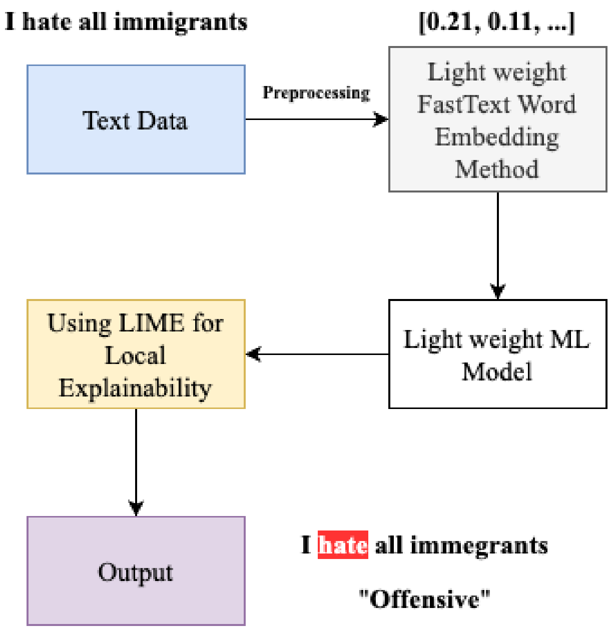


Figure 1: Visual Representation of our Problem Statement

3.3 Dataset Description

In this study, we utilized the HateXplain Dataset, which was sourced from the Kaggle repository [12]. Initially, the dataset comprised four features: Post_ID, Annotators, Post_Token, and Rationales. To determine the class label of each instance—categorized as 'hateful,' 'offensive,' or 'normal'—we relied on majority voting among the annotators. Instances lacking consensus, where the majority of annotators disagreed, were excluded to maintain data integrity. Additionally, we discarded any entries without a valid source. Following these refinements, the dataset was condensed to 19,201 valid instances, each characterized by three features: Post_ID, Post_Token, and label. In the Figure 2, 19,201 valid instances in the dataset, 7,800 are categorized as normal, 5,926 as hateful, and the remaining 5,475 instances are classified as offensive.

4 Methodology

This study implemented a rigorous five-step procedure to develop an explainable model for detecting abusive language, emphasizing both performance and interpretability. Each step is meticulously designed to address specific challenges associated with textual data analysis and machine learning.

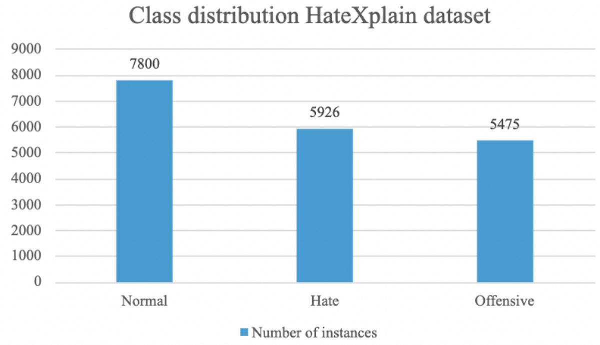


Figure 2: Class Distribution of HateXplain Dataset [12]

4.1 Data Sanitization

The initial phase involved comprehensive data sanitization. This included a series of preprocessing steps such as text normalization (converting text to lowercase, removing punctuation and numbers), tokenization, and removal of stopwords. These steps are crucial for reducing noise in the data and standardizing input formats, which helps improve the subsequent modeling stages.

4.2 Word Embedding with FastText

After preprocessing, we trained a FastText embedding model on the sanitized dataset to generate high-quality word embeddings. FastText is particularly adept at understanding the morphological nuances of words by using subword information, making it ideal for capturing the contextual subtleties needed for effective hate speech detection with limited memory usage and vocab size.

4.3 Handling Imbalanced Data with SMOTE

Recognizing the challenges posed by imbalanced datasets—a common issue in hate speech detection—we employed the Synthetic Minority Over-sampling Technique (SMOTE) [6]. This approach helps to balance the dataset by artificially generating synthetic samples from the minority class, thus enhancing the model's ability to learn from underrepresented classes.

4.4 Machine Learning Model

For the prediction of abusive language detection, four machine learning classifiers were employed, each known for their robustness in handling classification tasks. These classifiers include Decision Trees (DT), Extra Trees (ET), and Random Forest (RF). Detailed descriptions and the theoretical underpinnings of these well-known machine learning algorithms are extensively documented in the literature [3, 5, 16].

Subsequently, we utilized a stacking classifier (ST) approach to enhance predictive accuracy [18]. In this configuration, Decision Trees served as the final estimator, while Extra Trees and Random Forest acted as base estimators. This ensemble method allowed for leveraging the strengths of each individual classifier by using their predictions as input features to the final estimator. This approach effectively combines the diverse decision boundaries of each base estimator, thereby improving the overall predictive performance of the model.

4.5 Enhancing Robustness through Adversarial Testing

To further ensure the robustness of our model, we introduced adversarial testing into our methodology. This involved generating adversarial text samples through both word-level and character-level attacks [10]. These samples were then used to test and refine the model's ability to withstand malicious input designed to bypass hate speech detection.

4.6 Transparency with LIME

Finally, to address the need for transparency and explainability in automated decision-making, we integrated LIME into our workflow. LIME provides interpretable explanations for each prediction made by the model, highlighting the specific features or words that influenced the classification decision [17]. This is critical for validating the model's decisions and for maintaining user trust.

In addition to the methods described, our work involves a two-phase training process. Initially, we train our model using a dataset that does not include adversarial attacks, applying the aforementioned steps. Subsequently, we retrain the model using a dataset that has been augmented with adversarial attacks to compare and analyze the resilience and performance under these conditions.

5 Experimental Result

5.1 Experimental Set Up

In this study, 10-fold cross-validation was employed to enhance the robustness of the model and prevent overfitting. This method involves dividing the dataset into ten equal parts, with the model being trained on nine subsets and tested on the remaining one during each iteration [1]. This process was repeated ten times, each time with a different subset used as the test set, resulting in ten separate performance evaluations.

To ascertain the model's generalization ability, we computed the average performance across all iterations. The experiments were conducted using Google Colaboratory, leveraging Python packages such as SciKit Learn for model development, FastText for word embeddings, and LIME for model interpretability.

We utilized several metrics to evaluate the performance of our model, each highlighting different aspects of its effectiveness and reliability:

- **Accuracy**: Measures the proportion of true results (both true positives and true negatives) among the total number of cases examined.
- **Precision**: The ratio of true positive observations to the total predicted positives, indicating the accuracy of positive predictions.
- **Recall (Sensitivity)**: The ratio of true positive observations to the actual positives, measuring the model's ability to correctly identify all relevant instances.
- **F1 Score**: The harmonic mean of precision and recall, balancing both metrics for a binary classification model.

- **Cohen's Kappa**: Accounts for the agreement between two raters (or in predictive modeling, between the predictions and actuals) that is corrected for chance, providing a more robust measure than simple accuracy.
- **Matthews Correlation Coefficient (MCC)**: A correlation coefficient between the observed and predicted binary classifications, returning a value between -1 and +1, where +1 indicates a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement.

These metrics provide a comprehensive view of model performance, highlighting different aspects of its effectiveness and reliability.

5.2 Analyses and Findings

The table 1 provided offers a comprehensive performance comparison of various machine learning models under conditions of adversarial attacks and no attacks. The models evaluated include Decision Trees (DT), Random Forest (RF), Extra Trees (ET), and Stack Classifier (ST), across multiple performance metrics such as Accuracy, Precision, Recall, F1 Score, MCC (Matthews Correlation Coefficient), and Kappa.

5.2.1 ***Impact of Adversarial Attacks.*** Adversarial attacks generally decrease the performance of all models, although the extent varies by model. The DT model shows a significant reduction in accuracy from 66.01% in the absence of an attack to 64.78% under attack conditions. Conversely, ET and ST models exhibit a notable resilience, maintaining high performance levels even when attacked, suggesting a robustness against such manipulations.

5.2.2 ***Comparative Performance Analysis.*** In scenarios without adversarial interference, the RF and ET models outperform DT and ST in terms of accuracy, MCC, and Kappa, indicating better overall predictive capabilities. Notably, the F1 Scores of RF and ET are higher, reflecting their effective balance between precision and recall, crucial for managing both positive and negative classes efficiently.

5.2.3 ***Precision versus Recall Trade-off.*** A trade-off between precision and recall is evident among the models. RF and ET, while superior in accuracy, do not consistently lead in precision and recall, illustrating the compromises made between different types of errors.

5.2.4 ***Model Robustness to Adversarial Attacks.*** ET and ST models demonstrate minimal performance degradation when comparing conditions 'With Attack' and 'Without Attack', particularly in terms of MCC and Kappa. This resilience may be attributed to their ensemble methods, which provide multiple decision pathways, enhancing security against data manipulation.

This analysis reveals that ET and ST models, due to their minimal performance drop under adversarial attacks, might be more suitable in environments where resistance to adversarial interference is crucial. Meanwhile, the RF model, offering a robust balance across all metrics, is recommended for general applications where high accuracy and reliability are essential.

5.3 Local Interpretability using LIME

The figure 3 illustrates the output of an explainable machine learning model that classifies text data into three categories: Hateful, Normal, and Offensive. The left part of the figure displays the prediction probabilities assigned by the model for a specific text instance. These probabilities are:

- Hateful: 97%
- Normal: 1%
- Offensive: 2%

This distribution strongly suggests that the model has classified the text as Hateful with high confidence.

Adjacent to the prediction probabilities, the middle section of the figure provides a breakdown of the top contributing words to the classification decision, categorized by their impact on the classification as either increasing the likelihood of the text being Normal (right column in blue) or Not Normal (left column in orange). Words such as "niglets" and "rapist" have high weights, indicating a strong influence in pushing the classification towards Hateful, whereas other less weighted words contribute less significantly.

The far right section presents the text in question, with words highlighted according to their contribution to the model's decision. Words that highly influence the decision towards a Hateful classification are highlighted in red, such as "niglets" and "rapist," which are both pejorative and strongly associated with hate speech. In contrast, words like "white" and "voter," which appear less inflammatory, are highlighted in orange, indicating a lesser but still notable impact on the decision-making process.

This visualization serves a dual purpose:

Quantitative Analysis: It quantitatively shows the prediction confidence across different classes, emphasizing the model's decision-making process.

Qualitative Analysis: It qualitatively aids in understanding why certain classifications are made, highlighting specific words that influence the model's output. This level of insight is crucial for applications requiring transparency in automated content moderation, providing a basis for reviewing or contesting decisions made by AI.

By integrating LIME (Local Interpretable Model-agnostic Explanations), the system not only predicts but also explains its predictions, thereby enhancing the trustworthiness and accountability of automated content moderation systems.

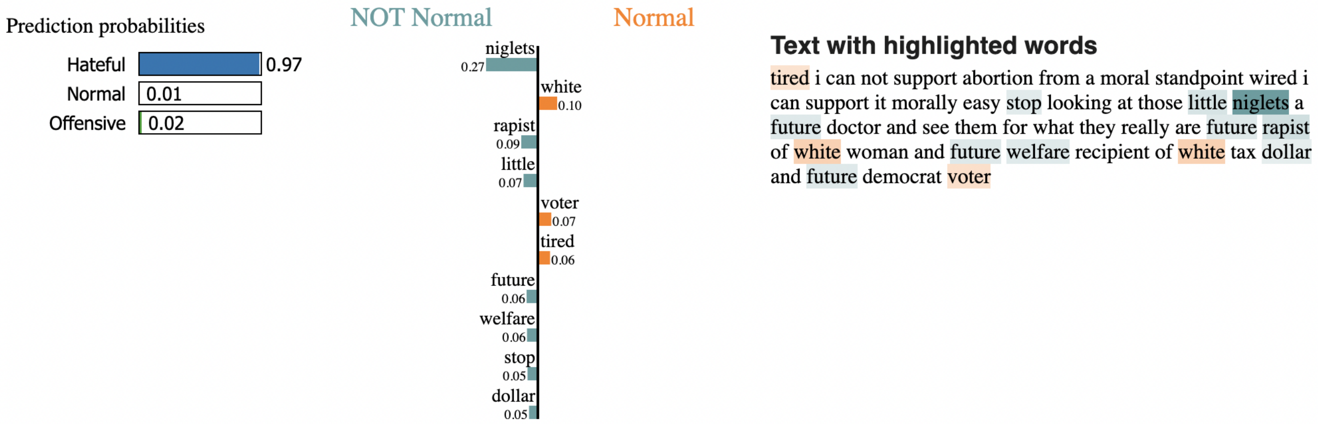
6 Conclusion & Future Work

6.1 Conclusion

In Table **??**, we compared our works from previous approaches. This research embarked on enhancing the interpretability and performance of machine learning models for text classification, focusing on the robust detection of nuances in textual data. Initially, the data was preprocessed to optimize it for analysis, followed by the utilization of FastText for advanced word embedding. A notable aspect of our study was the examination of the embedding size to fine-tune the balance between model complexity and performance efficiency.

Table 1: Performance Comparison of Models With and Without Adversarial Attacks

Model	Condition	Accuracy	Precision	Recall	F1 Score	MCC	Kappa
DT	With Attack	64.78%	64.93%	64.78%	64.63%	47.33%	47.17%
	Without Attack	66.01%	66.08%	66.02%	65.75%	49.25%	49.03%
RF	With Attack	73.17%	73.39%	73.17%	73.03%	59.98%	59.76%
	Without Attack	73.88%	74.35%	73.88%	73.66%	61.21%	60.84%
ET	With Attack	75.40%	75.71%	75.40%	75.22%	63.38%	63.10%
	Without Attack	75.41%%	76.09%	75.41%	75.16%	63.59%	63.11%
ST	With Attack	75.30%	75.98%	75.29%	75.13%	63.37%	62.94%
	Without Attack	74.64%	75.55%	74.64%	74.42%	62.53%	61.96%



tired i can not support abortion from a moral standpoint wired i can support it morally easy stop looking at those little niglets a future doctor and see them for what they really are future rapist of white woman and future welfare recipient of white tax dollar and future democrat voter

Figure 3: Impactness of word on individual predictions

Table 2: Comparison of Our Work with Previous Research

Metrics	**Previous Approach [12]**	**Our Approach**
Accuracy	69.8%	75.41%
Precision	-	76.09%
Recall	-	75.41%
F1-Score	68.7%	75.16%
MCC	-	63.59%
Kappa	-	63.11%
Memory Usage (RAM)	417.93 MiB	2.44 MiB

Our investigation led to the selection of the best performing lightweight model, which demonstrated significant efficacy in handling the tasks. However, when subjected to adversarial attacks, a slight decline in performance was observed, underscoring the challenges posed by sophisticated manipulation techniques.

To address the transparency and explainability of our model, we incorporated Local Interpretable Model-agnostic Explanations (LIME), which significantly aided in elucidating the decision-making processes of the model on a granular level. This approach not only bolstered the model's reliability but also its accountability in operational settings.

6.2 Future Work

Future research directions will primarily focus on enhancing the resilience of our models against adversarial attacks. More rigorous testing scenarios and diversified attack vectors will be employed to gauge the robustness of the model under more challenging conditions. Further optimization of FastText embedding dimensions may also be explored to reduce computational overhead while maintaining or enhancing model accuracy.

Another promising area of research involves the integration of additional explainable AI techniques alongside LIME to provide even deeper insights into model predictions and to foster greater trust among end-users. Moreover, extending our methodologies to other domains of text analysis, such as sentiment analysis or legal document review, could potentially validate the versatility and applicability of our approach.

Continued efforts will be made to refine these models, ensuring that they not only perform with high accuracy but also align with ethical AI standards, providing transparent and fair decision-making across various applications.

7 Contribution of this work

Name	ID$_Number$	% of Contribution
Soumik Datta	0424052089	50%
Anika Tasnim	0424058003	50%

References

[1] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. *Learning from data*. Vol. 4. Amlbook.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and machine learning. *arXiv preprint arXiv:1901.10439* (2019).

[3] Christopher M Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer Science+Business Media LLC., New York.

[4] Saugata Bose and Guoxin Su. 2022. Deep one-class hate speech detection model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 7040–7048.

[5] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[6] Nitesh Chawla, Kevin Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

[7] Margie Hertz. 2017. Online harassment 2017. *Pew Research Center* (2017).

[8] Fierce Inc. [n. d.]. Leading Business Problem 3: Lack of Transparency. https://fierceinc.com/leading-business-problem-3-lack-of-transparency/

[9] Joon Yul Kim, Robert O Wyatt, and Elihu Katz. 1999. Social connections in cyberspace. *Communication Research* 26, 6 (1999), 732–753.

[10] Shouling and Du Li, Jinfeng and Ji. [n. d.]. TextBugger: Generating Adversarial Text Against Real-world Applications. ([n. d.]).

[11] Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*. Springer, 415–426.

[12] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *AAAI Conference on Artificial Intelligence*.

[13] Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2022. BERT-based ensemble approaches for hate speech detection. In *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 4649–4654.

[14] Rachna Narula and Poonam Chaudhary. 2024. A comprehensive review on detection of hate speech for multi-lingual data. *Social Network Analysis and Mining* 14, 1 (2024), 1–35.

[15] Lam Nguyen Tung, Steven Cho, Xiaoning Du, Neelofar Neelofar, Valerio Terragni, Stefano Ruberto, and Aldeida Aleti. 2024. Automated Trustworthiness Oracle Generation for Machine Learning Text Classifiers. *arXiv e-prints* (2024), arXiv–2410.

[16] Geurts Pierre, Ernst Damien, and Wehenkel Louis. 2006. Extremely randomized trees. *Machine learning* 63, 1 (2006), 3–42.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[18] David H Wolpert. 1992. Stacked generalization. *Neural Networks* 5, 2 (1992), 241–259.