

DOI: https://doi.org/10.48009/3_iis_2026_131

Model agnostic explainable AI and ensemble machine learning approach for the cyberbullying classification

Piyush Vyas, *Texas A&M University-Central Texas*, piyush.vyas@tamuct.edu

Abstract

In the fast evolving age of artificial intelligence (AI), cyberbullying is becoming one of the emerging problems, specifically on social media platforms. Existing research has focused on resolving this issue utilizing machine learning (ML) and deep learning (DL) techniques. However, such solutions lack transparency, leaving the black box of machine learning and deep learning opaque and preventing understanding of how these techniques make decisions to detect or classify cyberbullying related content on social media. In this work, we address this gap by using an explainable AI (XAI) technique, Shapley Additive exPlanations (SHAP), along with an ensemble machine learning voting classifier combining random forests (RF) and XGBoost (XGB). In our experiment, with an accuracy of 71% and an area under the curve of 77.86%, the RF technique has performed well compared to the voting classifier and XGB. Further, we demonstrated the working of SHAP explanation on a sample text related to cyberbullying, utilizing the trained RF model with a combination of SHAP to show how the RF made its classification decision.

Keywords: Explainable AI, XAI, SHAP, Ensemble Learning, Cyberbullying

Introduction

As people's digital presence increases, so do the adverse challenges like cyberbullying, specifically on social media platforms like Facebook, Instagram, and X (formerly known as Twitter). According to Nagata et al. (2025), cyberbullying can be understood as the purposeful and ongoing mistreatment of a person through digital tools, including computers, mobile phones, and other electronic devices, often characterized by a power imbalance. Further, the authors reported an 8.7% lifetime prevalence among early adolescents in the USA, while Rostam-Abadi et al. (2024) found rates between 5.6% and 9.1% in high schools.

Beyond its online nature, cyberbullying is not only tied to the digital space (i.e., online environment), but its mental health effects are also severe; for example, the cyberbullying victims usually face a 2.62 times higher odds increase in suicidal behaviors (Nagata et al., 2025). Meta-analytic results by Li C et al. (2024) show that cyber-victims' odds ratio (i.e., metrics of likelihood) is 3.38 to experience depression and 3.52 to report suicidal ideation. Furthermore, Zhang S et al. (2025) indicate that specific online features exacerbate mental distress and substance use. These statistics underscore a significant public health crisis, where digital harassment significantly elevates risks for anxiety and self-harm across youth populations. Therefore, a systematic intervention approach remains crucial to mitigating these cumulative psychological outcomes created by cyberbullying. Thus, the objective of this research work is to classify the cyberbully tweet vs non-cyberbully tweet with inference utilizing an ensemble based classifier with model-agnostic XAI based SHAP technique.

Ensemble learning is a mechanism to build a powerful classifier combining multiple ML techniques to act as one unit. XAI provides a way to examine and explore the predictive ability of ML techniques, enabling interpretation and inference to open the black box nature of ML. SHAP is one of the XAI techniques usually used to provide the model agnostic (i.e., the technique is not restricted to a specific model type, and independently can provide explanations for individual samples) inference of text classification tasks like cyberbullying classification. To be specific, in our proposed approach, we used three classifiers, RF, XGB, and an ensemble voting combination of RF and XGB, and on top of that, whichever classifier performed well will be used by SHAP to explain the decision of the top performing classifier based on the performance accuracy.

Existing studies have explored the field of cyberbullying detection through diverse ML methodologies. For example, Unnava et al. (2024) and Kanitha et al. (2025) assessed Logistic Regression, Naïve Bayes, XGBoost, Decision Tree, and Random Forest as standalone models across large-scale cyberbullying datasets. Conversely, we are using the ensemble of ML techniques. Moreover, a relevant study, Alqahtani and Ilyas (2024), utilized ensemble-based detection for cyberbullying; however, they utilized the multi-class classification. Conversely, our work addresses binary classification on merged, deduplicated Kaggle datasets with full SHAP-based inference. Similarly, Mubeen M et al. (2025) used a stack ensemble approach for hate speech detection; however, it lacks the model-agnostic transparency to justify classifier predictions, as our approach does. Moreover, to unearth the ML/DL model black box in nature and on the interpretability aspect, Tesfagergish and Damaševičius (2024) and Mittal D et al. (2024) for XAI adoption for tasks like cyberbullying classification and hate speech. Thus, to complement such studies, our work uses XAI; however, their focus is on the interpretability of DL techniques, whereas ours focuses on ensemble techniques. Therefore, our work addresses this gap by pairing the ensemble technique with the XAI SHAP for binary cyberbullying classification.

Literature Review

Recently, many studies, such as Li Q et al. (2024) and Zhang S et al. (2025), have focused on understanding the impact of cyberbullying on individuals and society. Studies such as He Z et al. (2024) have also explored ways to solve this issue. Moreover, some studies, such as Unnava S and Parasana S (2024) and Mubeen M et al. (2025), have provided technical solutions to counter cyberbullying and its detection online.

Li Q et al. (2024) have studied the existing research on cyberbullying to understand the consequences of cyberbullying victimization across multiple disciplines. They categorized consequences into areas such as psychological, online behavioral, offline behavioral, offline interpersonal, and societal impacts. Zhang S et al. (2025) have examined the relationship between cyberbullying victimization, mental distress, and adolescent substance use. They identify publicity and separation as two important cyber features influencing how cyberbullying affects victims psychologically and behaviorally. In order to understand the further effects of cyberbullying, Au C (2025) has investigated the effects of cyberbullying on businesses, specifically focusing on online reviews. To be specific, they conducted a case study on the Farm House Restaurant to examine the damage of cyberbullying on business reputation and its impact on the owner's mental health. In the study by He Z et al. (2024), an experimental design was proposed to study AI-generated prompts that can encourage bystanders to report cyberbullying on social networking platforms. They investigated two dimensions of empathy: emotional recognition and emotional resonance through perspective-taking mechanisms. Also, Nguyen A et al. (2024) explored the unintended consequences of combining automated moderation bots with human moderation in online communities. They have theorized

three effects of moderation systems: emboldening effects within communities, disengagement among users, and spillover harassment into neighboring communities.

Furthermore, five major themes, such as age-appropriate design, theory-driven participation methods, and technology-user need alignment, can be identified in developing digital bullying and cyberbullying interventions involving children and educators (K. Chen & Chen, 2026). Also, educational tools, automated detection systems, and information systems (IS) designs for mitigation can be possible solutions (He et al., 2024). In line to provide the solution, the study by Unnava S and Parasana S (2024) has provided a comparative analysis of traditional ML/DL techniques for cyberbullying detection. They also show the role of feature engineering, preprocessing, and model selection in improving cyberbullying classification. Kanitha T et al. (2025) used several ML algorithms, such as Logistic Regression, Naive Bayes, XGB, Decision Trees, and RF, for cyberbullying detection. The study by Alqahtani and Ilyas (2024) develops ensemble-based ML approaches to classify six categories of cyberbullying content on social media. Here, the authors demonstrate the usefulness of ensemble learning for multi-class cyberbullying classification. Similarly, Mubeen M et al. (2025) have proposed a stacked ensemble framework for distinguishing hate speech, abusive language, and neutral content on social media. They highlighted the importance of contextual text understanding and ensemble learning in handling noisy and informal user-generated content. In the study by Gupta A et al. (2025), the authors proposed a Cross-Platform Cyberbullying Detection Algorithm (CPCD-Alg) that combines textual, behavioral, linguistic, and network-based features. Their framework avoids relying on username matching because users may change identities across platforms.

Moreover, XAI can improve transparency and interpretability in cyberbullying detection (Tesfagergish & Damaševičius, 2024) when combined with DL and interpretable feature engineering for cyberbullying prevention. For example, Mittal D et al. (2024) investigate XAI techniques such as SHAP, Local Interpretable Model-Agnostic Explanations (LIME), and Layer-wise Relevance Propagation (LRP) that can improve interpretability in hate speech detection models. They combined XAI with DL architectures to better explain how hate speech classification decisions are made.

Methodology

Figure 1 indicates the approach that we have discussed in detail in the sub-sections below.

Data Preparation- Collection and Preprocessing

We have collected the cyberbully tweet data from the Kaggle provided by Tomar S (2026) (link: <https://www.kaggle.com/datasets/soorajtomar/cyberbullying-tweets/data> , accessed: May 6, 2026). The dataset has a total of 11099 tweets labeled as 1 for Cyber Bullying, 0 for Not Cyber Bullying (non cyberbullying). The dataset is balanced, and duplicates have been removed. After collecting the data, we further performed the preprocessing task. First, we did the stop word removal, then we performed the lowercasing of each word. Second, we removed the URL, hashtags, non-alphabetic characters, and white spaces.

Features - TF-IDF

As we have used text data, vectorization becomes an essential task. For this, we have used Term Frequency–Inverse Document Frequency (TF-IDF), a feature-extraction technique commonly used for natural language processing tasks. TF-IDF generally indicates the importance of terms within a document relative to the entire data (Salton & Buckley, 1988).

Here, TF indicates the term frequency, which means how often a word appears in the document, and IDF indicates the inverse document frequency, which means how rare the word is across all documents. So, it usually assigns a lower weight to commonly occurring terms and a higher weight to the unique words. Due to its nature of converting the text into sparse numeric feature vectors prior to training the ML models, it is widely used (Ramos, 2003).

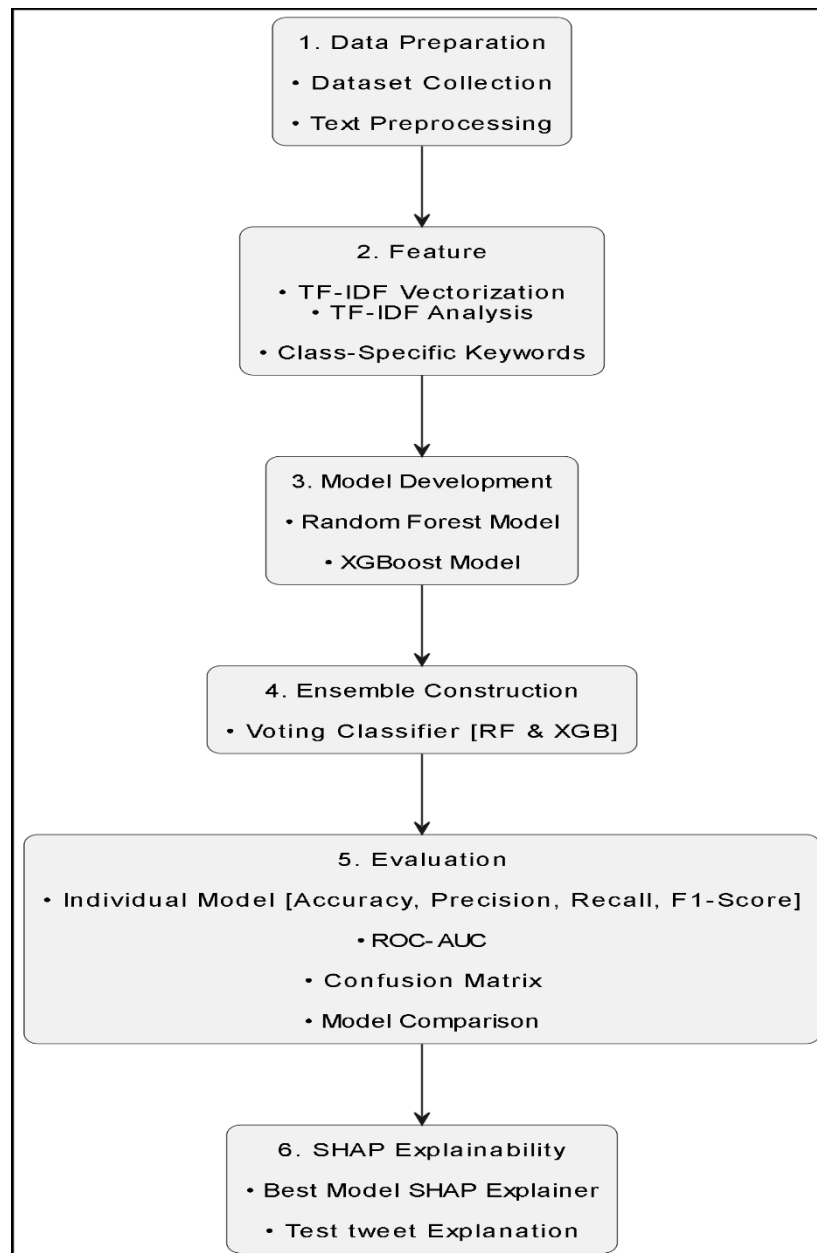


Figure 1. Overview illustration of the proposed approach.

Model Development: Classifiers

In this work, we have used ML techniques such as RF and XGB, both of which are tree-based classifiers. RF belongs to the ensemble learning model family because it builds many decision trees during the training.

This way it helps to reduce the overfitting by averaging the predictions from multiple trees trained on the subset of data (Breiman, 2001). Furthermore, XGB is based on the scalable gradient boosted tree mechanism. It sequentially builds the trees, and each new tree corrects the errors made by previous trees. This also helps to reduce the overfitting by using the regularization (T. Chen & Guestrin, 2016). Moreover, RF manages the high-dimensional sparse data like the TF-IDF, and requires minimal hyperparameter tuning (Breiman, 2001). Whereas XGB provides the built-in regularization, enables faster training via parallelization, and effectively handles class imbalance through weighted objectives (T. Chen & Guestrin, 2016).

Ensemble Voting Classifier

Ensemble learning is a mechanism of combining many base learners (i.e., ML techniques) to act as a single classification model that has the potential to outperform any single ML technique. The ensemble learning model reduces generalized errors by combining multiple ML models (Dietterich, 2000). Recent studies, such as Alghazzawi D et al. (2025), have asserted that ensemble techniques can enhance performance across diverse disciplines, including text classification. Moreover, aggregating predictions from heterogeneous base models as a meta-estimator is known as a voting classifier. where each model outputs class probabilities that are averaged, and the class with the highest average probability is selected (Pedregosa et al., 2011).

In our work, we combined the RF and XGB as a voting classifier. Here, both models independently generate probability estimates for each class. The voting classifier averages these probabilities with different weights, and the final prediction is the class with the highest average probability.

Evaluation

To evaluate the performance of the utilized techniques in this work, we have used accuracy, precision, recall, F1 score, and Receiver Operating Characteristic and Area Under the Curve (ROC-AUC) values. Here, accuracy indicates the performance of the model based on correctly classifying both classes. Precision indicates the proportion of predicted data samples of positive classes that are actually positive. Recall indicates the proportion of actual positive class samples that the model correctly identified. F1 score is the harmonic mean of precision and recall score, which provides a balanced measure and is useful for unbalanced data classification (Sokolova & Lapalme, 2009), ROC-AUC indicates the measurement of a model's ability to differentiate between classes among all possible classification thresholds, where 0.5 indicates random guessing, and 1.0 indicates perfect discrimination (Fawcett, 2006).

SHAP Explainability

XAI is the way to open up the black box behavior of ML techniques that help humans to understand and interpret the AI driven decisions made by these models (Barredo Arrieta et al., 2020). Its aim is to make models more transparent by using model-agnostic techniques such as SHAP. We can use advantages to increase model transparency, enhance debugging capabilities, and foster human AI collaboration. SHAP is an approach to model explainability that assigns each feature an importance value known as Shapley value. This value represents its marginal contribution to the prediction (Lundberg & Lee, 2017). SHAP is a frequently used method among other XAI techniques and works well with ML models and ensemble learning (Dietterich, 2000).

For tasks like text-based cyberbullying tweet classification, SHAP is particularly suitable because it helps identify which specific words or tokens push a prediction toward or away from a given class (e.g., cyberbullying or non-cyberbullying) (Lundberg et al., 2020). SHAP also produces intuitive visualizations (bar plot summaries) that communicate decision rationale to non-technical stakeholders.

Results and Discussion

As an experimental setup, to implement our approach, we have used Google Colab, an integrated cloud-based development environment. We have used a T4 GPU with 16GB VRAM and 12 GB system RAM. We have used Python 3 to code. Furthermore, for data cleaning, we have used Regex and NLTK Python-based libraries. We split the data in a stratified manner, with 75% of our data for training the ensemble techniques and 25% for testing the techniques. After splitting, our training set had 8325 posts, and the testing set had 2775 posts. More specifically, our training set class 0 had 4162 (50.0%) of posts and class 1 had 4163 (50.0%) posts. Our test set: class 0 had 1388 (50.0%) posts, and class 1 had 1387 (50.0%) posts. Our corpus vocabulary, or Dictionary size, is 8000. For hyperparameter tuning, we have used a 5-fold GridSearch cross-validation technique. The following Table 1 shows the tuned hyperparameters for the models and their descriptions. For RF, our best tuned parameters are max_depth = None, min_samples_leaf = 1, min_samples_split = 5, and n_estimators = 300. For XGB our best parameters are learning_rate = 0.1, max_depth = 9, n_estimators = 300, subsample = 0.8.

Table1. Hyperparameters of RF, XGB, and Voting classifier.

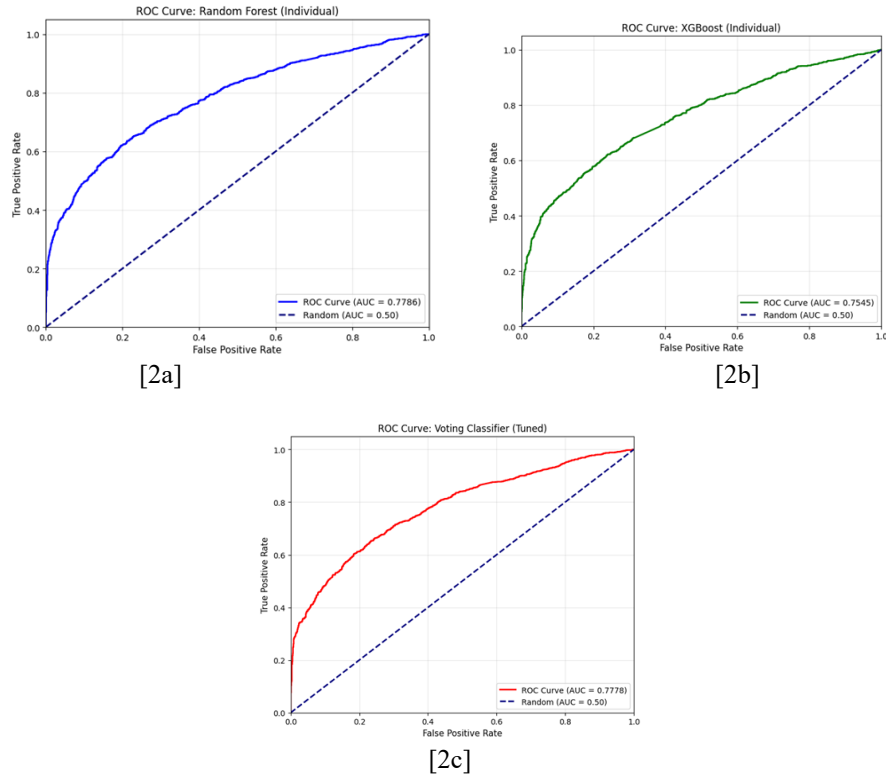
Techniques/ Classifiers	Parameters Range	Descriptions
RF	n_estimators = 100 - 300 max_depth = None and 10 - 30 min_samples_split = 2 & 5 min_samples_leaf = 1 & 2	- N_estimator is the total number of trees. - Max_depth is the maximum number of levels allowed in each decision tree. - Min_sample_split is the minimum number of samples required at a node before it splits into branches. - Min_sample_leaf is the minimum number of samples that must remain in the leaf node. - The learning rate is the step size used to control the errors of each new tree. - A subsample is a portion of the random training data selected for each tree. - Weights are assigned to each base classifier in an ensemble that determines how much influence each model's prediction has on the final combined output.
XGB	n_estimators = 100 - 300 max_depth = 3, 6, & 9 learning_rate = 0.01, 0.1, & 0.2 subsample = 0.8 & 1.0	
Ensemble Voting (RF + XGB)	Weights = [(1, 1), (1, 2), (2, 1)]	

Table 2. Top 10 words per class based on the TF-IDF mean.

Class 0 (non-cyberbullying)	(Cyberbullying)
Hate (0.029775)	F*ck (0.028421)
Damn (0.029259)	Hate (0.020100)
S*cks (0.019892)	As* (0.018727)
Lol (0.017274)	Damn (0.017371)
As* (0.017247)	Yours (0.017368)
Like (0.014199)	Like (0.017004)
f*ck (0.012869)	b*tch (0.015060)
love (0.011691)	idiot (0.014967)
would (0.011075)	f*cking (0.014841)
don't (0.010977)	s*cks (0.013544)

Furthermore, for the SHAP implementation, we have used a single sample text to demonstrate the best selected model's prediction ability utilizing the treexplainer function. Appendix Table 4 lists the libraries we used to build our approach. Table 2 shows the Top 10 Words per Class (by mean TF-IDF). Note that the Table 2 findings contain explicit, offensive, or derogatory language that we represent by single character replacement with a special symbol (*) for continuous reading, and at the same time, present the findings based on the collected dataset in nature.

Table 3. Evaluation measure score of each classifier.



Techniques	Precision %	Recall %	F1 Score %	Accuracy %	AUC
RF	72	71	71	71	77.86
XGB	69	69	68	69	75.45
Ensemble Voting (RF+XGB)	71	71	70	71	77.78

Figure 2. ROC-AUC curve of all three models (RF [2a], XGB [2b], and Voting classifier [2c]).

Table 3 shows the comparative analysis between RF, XGB, and the Ensemble voting classifier's evaluation measures. Wherein, RF performed better than the other two classifiers with 72% of precision, 71% of recall, 71% of F1 score, 77.86% AUC, and 71% of accuracy. Although these measures can be considered as moderately significant, the models have the potential to improve further. As RF performed better in this experiment, we further used it for the SHAP value explainability analysis of its prediction on the sample tweet text to demonstrate how RF makes a classification decision. Furthermore, Figures 2a,2b, and 2c show the ROC – AUC curve.

Figures 3a,3b, and 3c show the confusion matrix of all three classifiers. For RF, the true positive (TP) is 873, the true negative (TN) is 1096, the false positive (FP) is 292, and the false negative (FN) is 514, with a FP rate of 0.2104 and an FN rate of 0.3706. XGB TP is 883, TN is 1018, FP is 370, and FN is 504 with a FP rate of 0.2666 and a FN rate of 0.3634. For Voting Classifier (RF + XGB), TP is 863, TN is 1094, FP

is 294, and FN is 524 with FP Rate of 0.2118 and FN Rate of 0.3778. Overall, across these three classification metrics, it looks like RF has 29.05% (806/2775) misclassification, XGB has 31.50% (874/2775), and the Voting classifier has 29.87% (829/2775).

In Figures 4a, 4b, and 4c, we presented the probability distribution of each class for each of the classifiers used. The red color indicates class 0, and the blue color indicates class 1. The overlapping area indicates the misclassification or error (i.e., the model confuses between cyberbullying and non-cyberbullying text) in predicting the correct classes by each technique, for example, the XGB had a moderate misclassification rate from 0.2 to 0.8 majorly.

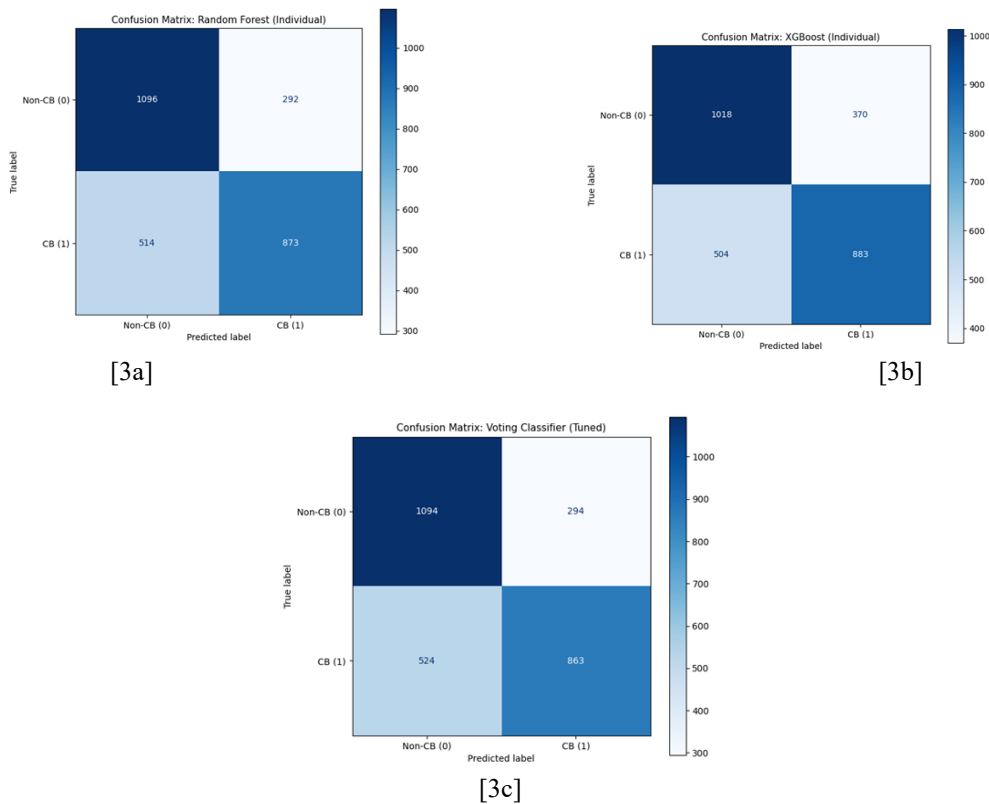


Figure 3. Confusion matrix of all three models (RF [3a], XGB [3b], and Voting classifier [3c]).

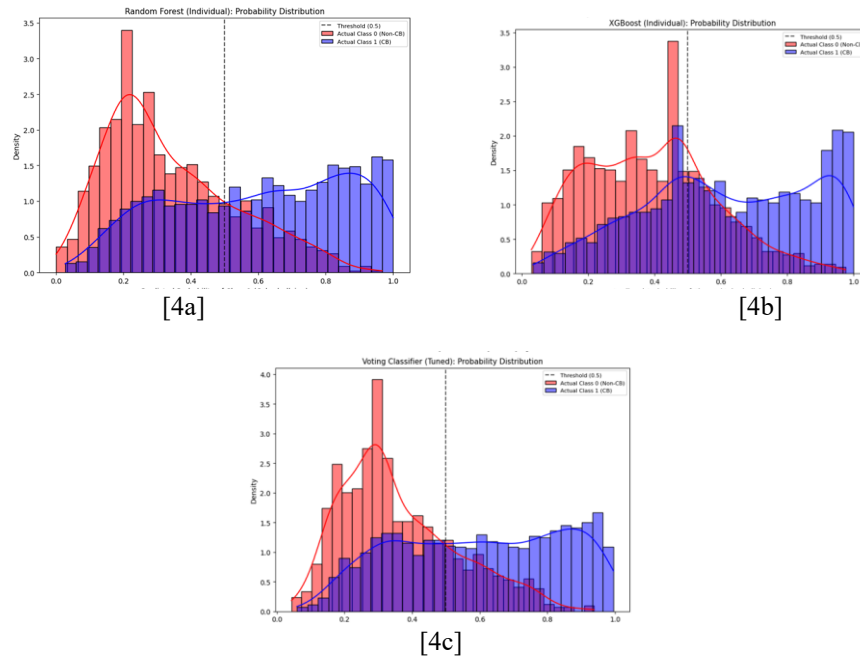


Figure 4. Classwise probability distribution of all three models (RF [4a], XGB [4b], and Voting classifier [4c]).

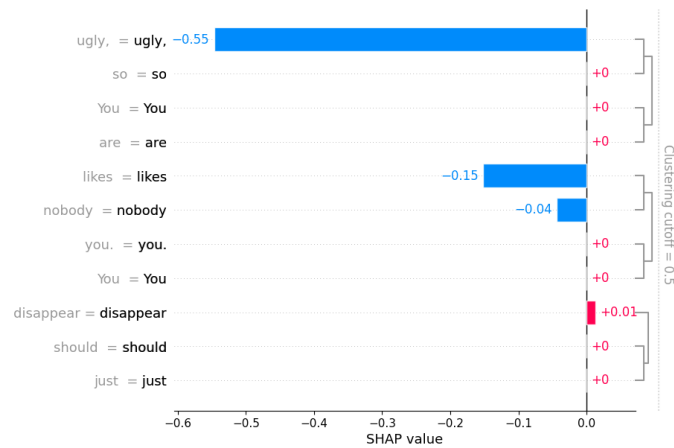


Figure 5. Class 0 (Not Cyberbully) word based SHAP bar plot.

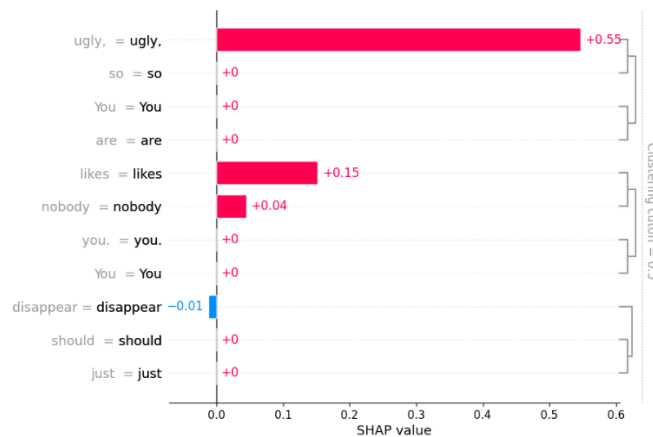


Figure 6. Class 1 (Cyberbully) word based SHAP bar plot.

Figures 5 and 6 demonstrate the SHAP working on a sample text like "You are so ugly, nobody likes you. You should just disappear." The RF has predicted it with confidence of being Class 0 (Non Cyberbullying)=0.0571, and being Class 1 (Cyberbullying)=0.9429. Then SHAP plotted the figure 5 and figure 6 bar graphs to show why RF made this decision by breaking each word in a text and showing its SHAP values. For example, in our sample text, as shown in Figure 5, the words like "ugly" have a SHAP value of -0.55; this negative value, represented as a blue bar pushed away from class 0 (non-cyberbullying), makes the model think that this text belongs to the other class (i.e., class 1). And in all the 0 values indicate no major contribution in the prediction of a specific class. SHAP values are the changing probability or marginal contribution value of a word from a base value (i.e., average value of all predictions) towards the final class prediction. Like in this case of class 0, the change is -55%. Conversely, if we look at Figure 6, the same word "ugly" leads the +55% marginal contribution shift from the base value towards class 1, indicating that because of this positive change, the model predicted the text as cyberbullying with 0.9429.

Implications for Management

From a managerial perspective, our work fosters trust and transparency in AI-based systems. Specifically, managers or policymakers can adopt XAI to make AI-derived decisions transparent to stakeholders, thereby reducing skepticism about black-box AI in content moderation. Moreover, XAI insights can guide the social media organization's platform policies on abrasive language patterns to support evidence-based guidelines, and the integration of explainable ML to detect cyberbullying may enhance reliability, given human validation. Moreover, we recommend that managers recognize AI limitations and maintain human-in-the-loop oversight in systems where evaluation measures are moderately high, like in our approach's results. Thus, continuous model improvement and dataset expansion should be prioritized.

Conclusion

As a potential contribution, our work presents a systematic approach for cyberbullying classification on social media, leveraging ensemble ML and model explainability. Also, this study positions the proposed approach as a replicable methodological pathway for researchers seeking to incorporate explainability into black-box classifiers, complementing and building upon existing literature while contributing toward the broader goal of transparent and trustworthy AI in the area of online safety. Moreover, future research can explore additional relevant datasets to improve the generalizability of the proposed approach, and advanced large language models can be used in combination with SHAP explanations. Also, involving contextual

embeddings to further minimize false negatives would be a plus. Moreover, our experiment can be extended by comparing our approach's results after further preprocessing with stemming and lemmatization.

Acknowledgements

We acknowledge all entities (affiliated university, reviewers, editors, experts) for all direct and indirect support. We have used the Grammarly MS Word plugin and Claude AI for writing assistance.

References

- Alghazzawi, D., Ullah, H., Tabassum, N., Badri, S. K., & Asghar, M. Z. (2025). Explainable AI-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. *Scientific Reports*, 15(1), 1111.
- Alqahtani, A. F., & Ilyas, M. (2024). An ensemble-based multi-classification machine learning classifiers approach to detect multiple classes of cyberbullying. *Machine Learning and Knowledge Extraction*, 6(1), 156–170.
- Au, C. H. (2025). The Impact of Cyberbullying on Businesses: An Affordance and Constraint Perspective. *PACIS 2025 Proceedings*. 7.
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, K., & Chen, Q. (2026). Participatory Design in Digital Interventions for Bullying and Cyberbullying: A Systematic Review. *International Journal of Human–Computer Interaction*, 0(0), 1–17. <https://doi.org/10.1080/10447318.2026.2632150>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1–15.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Gupta, A., Matta, P., & Pant, B. (2025). An integrated user matching framework for cross-platform cyberbullying detection. *Discover Computing*, 28(1), 224. <https://doi.org/10.1007/s10791-025-09743-7>
- He, Z., Li, Y.-J., & Lee, M. K. O. (2024). IT Solutions for Tackling Cyberbullying: A Literature Review. *PACIS 2024 Proceedings*. 6.

- Li, C., Wang, P., Martin-Moratinos, M., Bella-Fernández, M., & Blasco-Fontecilla, H. (2024). Traditional bullying and cyberbullying in the digital age and its associated mental health problems in children and adolescents: a meta-analysis. *European Child & Adolescent Psychiatry*, 33(9), 2895–2909. <https://doi.org/10.1007/s00787-022-02128-x>
- Li, Q., Liu, L., & Turel, O. (2024). Uncovering the Ripple Effect of Cyberbullying Victimization on Social Media. *ICIS 2024 Proceedings*. 10.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, 4768–4777.
- Mittal, D., Singh, H., & Rani, S. (2024). Exploring Explainable Artificial Intelligence Techniques for Hate Speech Detection. In A. E. Hassanien, S. Anand, A. Jaiswal, & P. Kumar (Eds.), *Innovative Computing and Communications* (pp. 545–554). Springer Nature Singapore.
- Mubeen, M., Muskan, A., Akram, A., Rashid, J., Alshalali, T. A. N., & Sarwar, N. (2025). Cyberbullying-Related Automated Hate Speech Detection on Social Media Platforms Using Stack Ensemble Classification Method. *International Journal of Computational Intelligence Systems*, 18(1), 174. <https://doi.org/10.1007/s44196-025-00919-z>
- Nagata, J. M., Shim, J., Balasubramanian, P., Leong, A. W., Smith-Russack, Z., Shao, I. Y., Al-Shoaibi, A. A. A., Helmer, C. K., Ganson, K. T., Testa, A., Kiss, O., He, J., Groves, A. K., Baird, S., & Baker, F. C. (2025). Cyberbullying, mental health, and substance use experimentation among early adolescents: a prospective cohort study. *The Lancet Regional Health – Americas*, 46. <https://doi.org/10.1016/j.lana.2025.101002>
- Nguyen, A., Rai, A., & Maruping, L. (2024). Understanding the Unintended Effects of Human-Machine Moderation in Addressing Harassment within Online Communities. *Journal of Management Information Systems*, 41(2), 341–366. <https://doi.org/10.1080/07421222.2024.2340831>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12, 2825–2830.
- Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*, 242(1), 29–48.
- Rostam-Abadi, Y., Stefanovics, E. A., Zhai, Z. W., & Potenza, M. N. (2024). An exploratory study of the prevalence and adverse associations of in-school traditional bullying and cyberbullying among adolescents in Connecticut. *Journal of Psychiatric Research*, 173, 372–380. <https://doi.org/10.1016/j.jpsychires.2024.03.033>
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5), 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0)

- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inf. Process. Manage.*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- T, K., K.R., D., & C., K. (2025). Cyberbullying Detection using Machine Learning Models. *Journal of Information Technology and Digital World*, 6(4), 362–372. <https://doi.org/10.36548/jitdw.2024.4.004>
- Tesfagergish, S. G., & Damaševičius, R. (2024). Explainable Artificial Intelligence for Combating Cyberbullying. In K. K. Patel, K. C. Santosh, A. Patel, & A. Ghosh (Eds.), *Soft Computing and Its Engineering Applications* (pp. 54–67). Springer Nature Switzerland.
- Tomar, S. (2026). *Cyberbullying Tweets*. <https://www.kaggle.com/datasets/soorajtomar/cyberbullying-tweets/data>.
- Unnava, S., & Parasana, S. R. (2024). A study of cyberbullying detection and classification techniques: A machine learning approach. *Engineering, Technology & Applied Science Research*, 14(4), 15607–15613.
- Zhang, S., Hao, Z., Leidner, D. E., & Cowan, B. W. (2025). Cyberbullying Victimization, Mental Distress, and Adolescent Substance Use: The Role of Cyber Features. *Information Systems Research*. <https://doi.org/10.1287/isre.2022.0454>

Appendix

Table 4. Utilized key Python libraries and their usage

Library	Purpose
pandas	Data loading
re	Regex-based text cleaning
nltk	English stopwords corpus and removal
scikit-learn	TF-IDF, train test split, GridSearchCV, classification metrics, confusion matrix, ROC/AUC
xgboost	XGBoost - gradient boosted tree classifier
shap	Model explainability (TreeExplainer, bar plots)
matplotlib	Visualization- ROC curves, bar charts
seaborn	Visualization - Confusion matrix heatmaps, probability distribution plots