Evaluating Trustworthiness and Interpretability of eXplainable AI Techniques for Cyberbullying Detection

Giuseppe Loseto[1], Filippo Gramegna[1], Agnese Pinto[2], Floriano Scioscia[2] and Michele Ruta[2]

[1]LUM "Giuseppe Degennaro" University, Department of Engineering, S.S. 100 km 18, Casamassima, I-70010, Italy
[2]Polytechnic University of Bari, Department of Electrical and Information Engineering, Bari, Via E. Orabona 4, I-70126, Italy

Abstract

Explainable Artificial Intelligence (XAI) has become a key research area for enhancing the trustworthiness of AI-based applications. In Cyber Social Security (CSS) contexts, transparent models and interpretable explanations are essential to support responsible integration into digital safety workflows. This work investigates the application of two complementary XAI techniques, LLM-based explanations and SHAP value analysis, to improve the trustworthiness of Machine Learning models for cyberbullying detection. An early evaluation involving 100 participants assesses how different user groups perceive the clarity and usefulness of these explanations. The findings highlight the benefits of multi-layered XAI strategies in supporting user understanding and confidence in automated moderation tools.

Keywords

eXplainable Artificial Intelligence, Cyberbullying Detection, Large Language Models, SHAP

1. Introduction and Motivation

Online social platforms have profoundly transformed interpersonal communication, enabling unprecedented opportunities for connection, collaboration, and self-expression [1]. However, this digital ecosystem has also amplified the prevalence of harmful behaviors, among which cyberbullying remains one of the most pervasive and damaging [2]. Manifesting through insults, threats, harassment, and other forms of abusive discourse, cyberbullying poses significant psychological, social, and even physical risks to individuals, particularly adolescents and vulnerable categories [3]. Traditional manual moderation approaches have become insufficient, motivating increased reliance on Artificial Intelligence (AI)-driven cyberbullying detection [4, 5]. Nevertheless, current systems operate as complex opaque boxes, making it difficult for both platform moderators and end-users to understand why a particular message, post or comment was flagged. This lack of transparency undermines the trustworthiness of automated moderation tools. To address these concerns, eXplainable Artificial Intelligence (XAI) has emerged as a crucial research area aimed at improving the interpretability of AI systems deployed in several scenarios, including healthcare [6], home automation [7], multi-agent systems [8], and Cyber Social Security (CSS) contexts [9]. XAI techniques seek to make the internal logic of complex models understandable by providing insights into why a certain prediction was made and which factors influenced the decision.

This research investigates the use of two XAI techniques, respectively based on Large Language Models (LLMs) and SHapley Additive exPlanations (SHAP), [10] to improve the trustworthiness of cyberbullying detection models, also proposing an early evaluation of their effectiveness. The paper is organized as follows: Section 2 introduces the proposed cyberbullying detection architecture; Section 3

1st Workshop on Symbiotic AI for Cyber Social Security: Human–AI Collaboration for Social Resilience and Trustworthy Digital Ecosystems (SAI4CCS 2026), co-located with the 2nd Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09–13, 2026, Cagliari, Italy

loseto@lum.it (G. Loseto); gramegna@lum.it (F. Gramegna); agnese.pinto@poliba.it (A. Pinto); floriano.scioscia@poliba.it (F. Scioscia); michele.ruta@poliba.it (M. Ruta)
https://www.lum.it/giuseppe-loseto/ (G. Loseto); https://www.lum.it/filippo-gramegna/ (F. Gramegna); https://sisinflab.poliba.it/pinto/ (A. Pinto); https://sisinflab.poliba.it/scioscia/ (F. Scioscia); https://sisinflab.poliba.it/ruta/ (M. Ruta)
0000-0002-7995-8494 (G. Loseto); 0000-0003-4162-957X (F. Gramegna); 0000-0002-2502-7467 (A. Pinto); 0000-0002-7859-9602 (F. Scioscia); 0000-0003-2125-327X (M. Ruta)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-14

CEUR-WS.org/Vol-4329/paper12.pdf

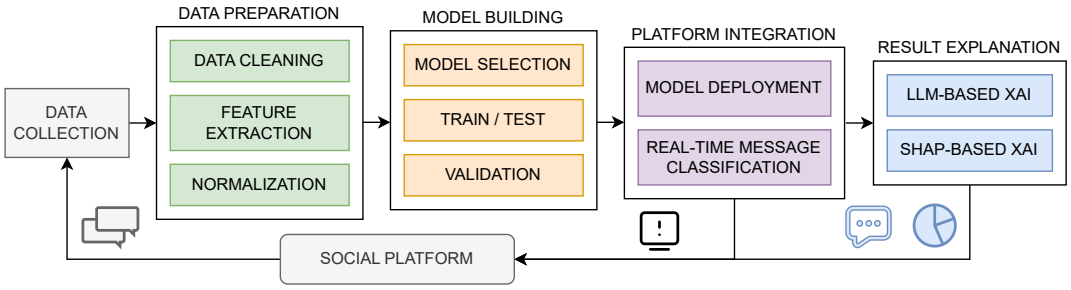


Figure 1: Reference architecture for explainable message classification

presents findings from the explainability study involving a group of 100 participants; finally, Section 4 provides directions for future work.

2. Cyber Social Security Platform

The proposed framework (Figure 1) is designed to detect cyberbullying in online social platforms through an integrated architecture combining Machine Learning (ML) classification models with two different XAI techniques. The system follows a modular design modeling a complete processing pipeline consisting of data preparation, model building and deployment and a human-in-the-loop XAI module supporting user awareness of classification results. The data collection module retrieves posts, information about user profiles, and contextual metadata from online platforms. Following the dataset proposed in [11], the feature extraction stage relies on a heterogeneous set of input attributes, reported in Table 1, that capture complementary dimensions of online interactions and model the broader social and communicative context in which they occur. Message-level features describe the behavioral tendencies and interaction patterns of users with respect to a specific message, also capturing aspects such as message propagation and activity patterns within the online community. User-level features characterize the degree of user engagement, the activity history on the platform, and structural aspects of social relationships. Linguistic and stylistic features capture the textual structure and expressive characteristics of user communications in order to characterize message composition and writing style.

The framework integrates different ML classification models, implemented using the *scikit-learn* library (https://scikit-learn.org/, v1.7.2), to ensure flexibility and adaptability to different cyberbullying detection scenarios. Specifically, the system includes Support Vector Machines (SVM), Logistic Regression (LR), k-Nearest Neighbors (k-NN), Decision Trees (DT), Random Forests (RF), and Gradient Boosting (GB). Any of these classifiers can be selected as the primary reference model depending on the computational constraints and the desired interpretability level. Classification models have been integrated into an web dashboard, developed using *Streamlit* (https://streamlit.io/, v1.51.0), presenting a chronological timeline of messages posted by different users. The model automatically evaluates the message and determines whether it exhibits characteristics associated with cyberbullying. As an illustrative

Table 1
Social features extracted for message classification

Category	Features
Message-level	IsRetweet (boolean), IsSelfMentioned (boolean), Num. of Retweets, Num. of Favorites, Num. of Hashtags, Num. of Medias, Num. of Mentions
User-level	Age of Sender Account in years, Num. of Sender Favorites, Num. of Sender Followings, Num. of Sender Followers, Num. of Sender Statuses, Sender Location
Linguistic and Stylistic	Num. of Emojis, Num. of Punctuations, Num. of Uppercase Letters, Num. of Letters, Num. of Symbols, Num. of Words, Num. of Stop Words, Num. of Unique Words, Num. of Slang Words, Average Word Length

Table 2
Co-12 Explanation Quality Properties

Content	Correctness (C1), Completeness (C2), Consistency (C3), Continuity (C4), Contrastivity (C5), Covariate complexity (C6)
Presentation	Compactness (C7), Composition (C8), Confidence (C9)
User	Context-dependence (C10), Coherence (C11), Controllability (C12)

example, we consider a harmful message posted by *@user1207* to offend another individual on the social network. When a message is flagged as potentially harmful, the system adopts a protective disclosure mechanism in which the original content is temporarily replaced by a warning notification informing the user about the suspicious nature of the message. The user may then choose to reveal the hidden post, thereby maintaining both transparency and user agency while encouraging safer interaction patterns within the platform. The dashboard provides a dedicated explanation section (Figure 2) presenting a textual explanation, generated by LLaMA 3.2 3B model [12], offering a human-readable summary of the factors influencing the classification. Users may also open an additional analysis-oriented panel that provides three SHAP-based visualizations: (i) a *beeswarm* chart illustrating the contribution of individual features, e.g., user profile attributes or stylistic attributes, to the overall classification model; (ii) a *waterfall* plot reporting a step-by-step and ordered decomposition of the explanation for a single prediction; (iii) a *force* plot showing how specific attribute contributions push the prediction higher or lower, using a more fluid and intuitive visual representation. Due to space constraints, Figure 3 presents, as a representative example, the force plot corresponding to the same cyberbullying message described above. Features with positive (red) contributions shift the prediction toward cyberbullying, whereas features with negative (blue) contributions move it toward non-cyberbullying.

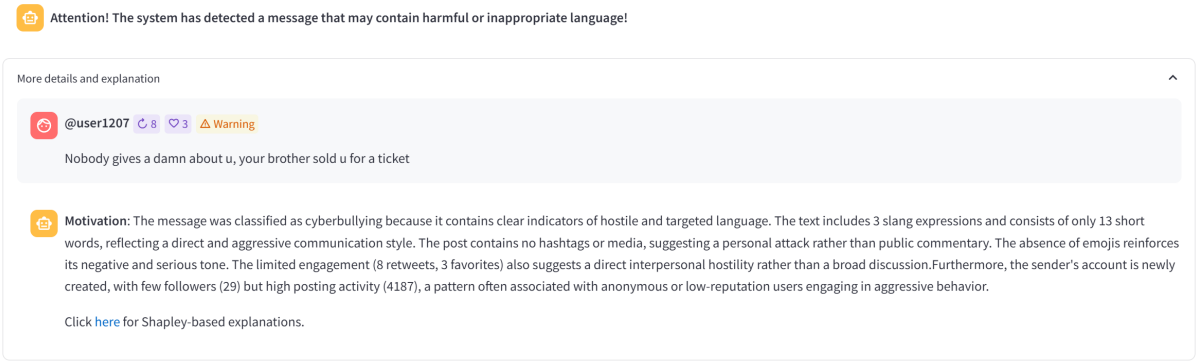


Figure 2: LLM-based textual explanation

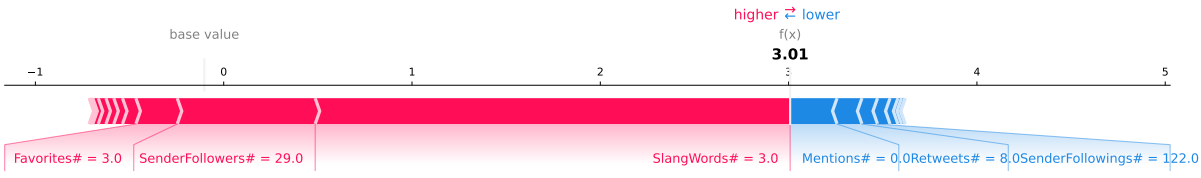
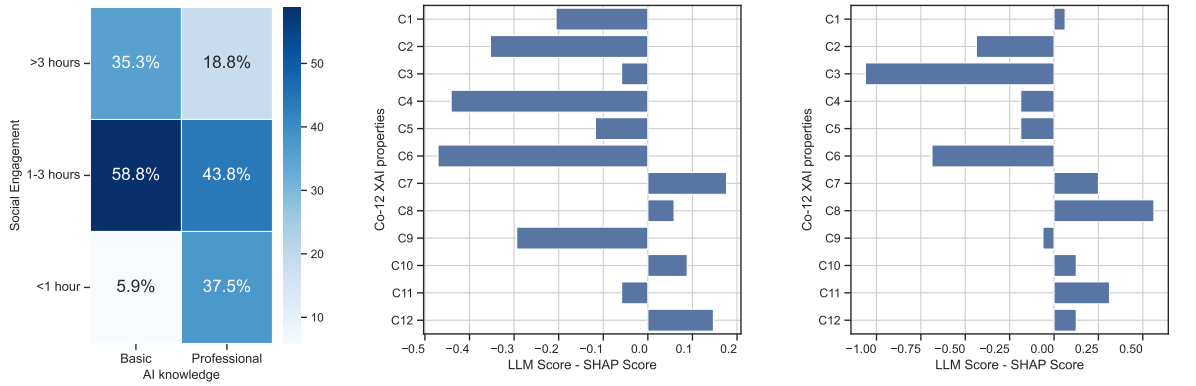


Figure 3: SHAP-based force plot

3. Evaluation of eXplainable AI Techniques

The assessment of the two proposed XAI techniques has been conducted using the *Co-12* explanation quality properties [13] as reference criteria for evaluating user perceptions of explanation format and

structure, with particular emphasis on how classification outcomes are communicated and justified. These properties enable a systematic comparison according to three different explanation dimensions (content, presentation, user) detailed in Table 2. A survey has been administered to 100 participants exhibiting diverse backgrounds in terms of familiarity with artificial intelligence and level of engagement with social platforms (Figure 4a). Particularly, the sample comprised 68 *basic* users with simple AI skills and 32 *professional* users with advanced AI knowledge, who exhibited distinct evaluation patterns. A direct comparison of these differences is shown in Figure 4b and Figure 4c, where positive values indicate a preference for LLM-based explanations, while negative values reflect a preference for SHAP. Basic users rate SHAP as slightly more informative for content-based dimensions (C1-C6), but consider LLM explanations more user-oriented due to their compactness (C7) and better alignment with the users' context and expertise (C10). SHAP, however, receives higher confidence scores (C9) thanks to its precise, probability-based indicators. Professional users show a larger gap between SHAP and LLM-based explanations. They strongly prefer SHAP for content-related properties (C2, C3, C6), valuing its quantitative depth, visual clarity, and feature-level specificity. However, they prefer the presentation format and structural organization of LLM-based explanations (C8), as well as their coherence with users' background knowledge, beliefs, and general consensus (C11).



(a) Users distribution (b) XAI evaluation (basic users) (c) XAI evaluation (professional users)

Figure 4: Evaluation results grouped by user profiles

4. Conclusion

This study introduced a cyberbullying detection framework combining ML models with two complementary XAI techniques to improve user interpretability of classification results. The evaluation shows that SHAP explanations are perceived as more detailed and informative, while LLM-generated outputs are appreciated for their clarity and user-friendly presentation, highlighting the benefits of a multi-layered explainability strategy tailored to users needs and technical backgrounds. Future work will expand the XAI framework by incorporating real-time user feedback, exploring additional LLMs for richer interpretative outputs, and integrating authoritative response guidelines and regulations to provide users with actionable advice when cyberbullying is detected.

Acknowledgments

This work has been supported by the project titled "Cyber Social Security" (CSS) - National Recovery and Resilience Plan, Mission 4 "Education and Research" – Component 2 "From Research to Business" – Investment 1.3, funded by the European Union – NextGenerationEU – CUP F53C22000740007.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Alsaleh, The impact of technological advancement on culture and society, Scientific Reports 14 (2024).

[2] M. Mladenović, V. Ošmjanski, S. V. Stanković, Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges, ACM Computing Surveys (CSUR) 54 (2021) 1–42.

[3] G. Gohal, A. Alqassim, E. Eltyeb, A. Rayyani, B. Hakami, A. Al Faqih, A. Hakami, A. Qadri, M. Mahfouz, Prevalence and related risks of cyberbullying and its effects on adolescent, BMC psychiatry 23 (2023) 39.

[4] L. B. Varra, T. M. Singh, K. Lippert, et al., AI-Based Cyberbullying Detection Techniques and Strategies, in: Combating Cyberbullying With Generative AI, IGI Global Scientific Publishing, 2025, pp. 63–94.

[5] S. Wang, A. S. Shibghatullah, T. J. Iqbal, K. H. Keoy, A review of multimodal-based emotion recognition techniques for cyberbullying detection in online social media platforms, Neural Computing and Applications 36 (2024) 21923–21956.

[6] W. V. Woensel, F. Scioscia, G. Loseto, O. Seneviratne, E. Patton, S. Abidi, L. Kagal, Explainable Clinical Decision Support: Towards Patient-Facing Explanations for Education and Long-Term Behavior Change, in: M. Michalowski, S. S. R. Abidi, S. Abidi (Eds.), Artificial Intelligence in Medicine, Springer International Publishing, Cham, 2022, pp. 57–62.

[7] M. Ruta, F. Scioscia, G. Loseto, E. Di Sciascio, A Semantic-Enabled Social Network of Devices for Building Automation, IEEE Transactions on Industrial Informatics 13 (2017) 3379–3388.

[8] C. Fasciano, G. Loseto, A. Pinto, M. Ruta, F. Scioscia, A propagation-based ranking semantics in Explainable Bipolar Weighted Argumentation, Engineering Applications of Artificial Intelligence 142 (2025) 109767.

[9] V. S. Barletta, D. Caivano, C. Catalano, S. del Vescovo, M. Scalera, Towards the Responsible/Trustworthy AI in Multi-Domain Operations for Cyber Social Security: A Black-Box AML Case Study in the CAN Bus Frame Detection Task, in: G. Costa, R. Montanari, M. Carminati, G. Sciarretta (Eds.), Proceedings of the Joint National Conference on Cybersecurity (ITASEC & SERICS 2025), Bologna, Italy, February 03-08, 2025, volume 3962 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[10] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems (NeurIPS) 30 (2017).

[11] A. Bozyiğit, S. Utku, E. Nasibov, Cyberbullying detection: Utilizing social media features, Expert Systems with Applications 179 (2021) 115001.

[12] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv e-prints (2024) arXiv–2407.

[13] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, C. Seifert, From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI, ACM Computing Surveys 55 (2023) 1–42.