Explainable Artificial Intelligence for Combating Cyberbullying

Senait Gebremichael Tesfagergish[1] and Robertas Damaševičius[1,2](✉)

[1] Department of Software Engineering, Kaunas University of Technology,
Kaunas, Lithuania
sengeb@ktu.lt

[2] Department of Applied Informatics, Vytautas Magnus University,
Kaunas, Lithuania
robertas.damasevicius@vdu.lt

Abstract. Cyberbullying has become a serious societal issue that affects millions of people globally, particularly the younger generation. Although existing machine learning and artificial intelligence (AI) methods for detecting and stopping cyberbullying have showed promise, their interpretability, reliability, and adoption by stakeholders are sometimes constrained by their black-box nature. This study presents a thorough description of Explainable Artificial Intelligence (XAI), which tries to close the gap between AI's strength and its interpretability, for preventing cyberbullying. The first section of the study examines the prevalence of cyberbullying today, its effects, and the limits of current AI-based detection techniques. Then, we introduce XAI, outlining its significance and going through several XAI frameworks and methodologies. The major emphasis is on the use of XAI in cyberbullying detection and prevention, which includes explainable deep learning models, interpretable feature engineering, and hybrid methods. We examine the ethical and privacy issues surrounding XAI in this setting while presenting many case cases. Additionally, we compare XAI-driven models with conventional AI techniques and give an overview of assessment criteria and datasets relevant to the identification of cyberbullying. The article also examines real-time alarm systems, defensible moderator suggestions, and user-specific feedback as XAI-driven cyberbullying intervention tools. We conclude by outlining possible future developments, such as technology advancements, fusion with other developing technologies, and resolving difficulties with prejudice and fairness.

Keywords: Explainable Artificial Intelligence · Cyberbullying · Machine Learning · Deep Learning · Text Analysis · Interpretability

1 Introduction

In today's digital era, the Internet and social media platforms have become an integral part of people's lives, fostering communication, information sharing, and collaboration. However, these advancements have also given rise to a darker side of human interaction, known as cyberbullying. Cyberbullying refers to the

K. K. Patel et al. (Eds.): icSoftComp 2023, CCIS 2030, pp. 54–67, 2024.
https://doi.org/10.1007/978-3-031-53731-8_5

use of digital technologies to harass, threaten, or harm others deliberately and repeatedly. It has emerged as a significant social problem, impacting the mental health and well-being of millions of individuals worldwide, especially among the younger population.

Traditional Artificial Intelligence (AI) and Machine Learning (ML) approaches have been utilized to detect and prevent cyberbullying, employing techniques such as keyword-based filtering, supervised machine learning, and deep learning models. While these methods have shown promise in identifying and addressing cyberbullying incidents, they often suffer from a lack of interpretability. This black-box nature of AI algorithms raises concerns regarding their trustworthiness, accountability, and acceptance by various stakeholders, including end-users, platform moderators, and policymakers.

Explainable Artificial Intelligence (XAI) has emerged as a potential solution to address these challenges, offering insights into the decision-making process of AI models while maintaining high performance [15]. XAI techniques facilitate the understanding of complex AI models by providing human-readable explanations, thereby enhancing trust, transparency, and control over automated systems.

The motivation behind this paper is to present a comprehensive overview of XAI for combating cyberbullying, exploring its potential benefits and challenges in the context of cyberbullying detection and prevention. By examining XAI-driven approaches and their applications, this paper aims to foster a better understanding of XAI's role in this domain and contribute to the development of more effective, transparent, and trustworthy AI-driven solutions to tackle the pressing issue of cyberbullying.

Cyberbullying is a pervasive and detrimental social issue affecting millions of individuals worldwide, particularly among the younger population. Traditional AI and ML techniques have demonstrated potential in detecting and preventing cyberbullying. However, these methods often suffer from a lack of interpretability, raising concerns about their trustworthiness, accountability, and acceptance by various stakeholders. The primary challenge is to develop and implement XAI-driven approaches that provide clear, human-readable explanations for their decisions while maintaining high performance in detecting and preventing cyberbullying incidents.

2 Artificial Intelligence and Machine Learning in Cyberbullying Detection

Traditional approaches for cyberbullying detection primarily rely on keyword-based filtering, supervised machine learning, and deep learning models [24]. Keyword-based filtering is a relatively simple method that involves identifying and blocking messages containing explicit or offensive words and phrases. However, this approach has limited effectiveness as it is often unable to capture the nuances and subtleties of natural language, leading to high false positive and false negative rates. Moreover, cyberbullies may deliberately misspell words or use euphemisms to evade detection, further undermining the efficacy of keyword-based filtering.

Supervised machine learning methods have been employed to address some of these limitations. These approaches involve training classifiers on labeled datasets comprising instances of cyberbullying and non-cyberbullying content. Commonly used algorithms include Support Vector Machines (SVM), Naïve Bayes, and Decision Trees, among others. These methods analyze textual features, such as term frequency-inverse document frequency (TF-IDF) and n-grams, to identify patterns associated with cyberbullying. Although supervised machine learning techniques have demonstrated improvements over keyword-based filtering, they still struggle with the complexities and ambiguities of natural language, and their performance can be highly dependent on the quality and representativeness of the training data.

Deep learning models, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have also been applied to cyberbullying detection. These models are capable of capturing more complex patterns in textual data, as they automatically learn relevant features through multiple layers of representation. While deep learning methods have shown promising results in some cases, they are computationally intensive and often require large amounts of labeled data for training. Furthermore, these models are typically considered black-box algorithms, as their decision-making processes are difficult to interpret and understand.

Related works focus on the application of machine learning and deep learning techniques for detecting and mitigating abusive language and hate speech on social media platforms [3]. The papers also emphasize the importance of explainability in the decision-making process of these models. Wich et al. [25], proposes a method for detecting abusive language on social media platforms using user and network data. The authors use a combination of shallow and deep learning algorithms and evaluate their performance in terms of accuracy and recall. The results show that bidirectional long-short-term memory (LSTM) is the most efficient method for detecting abusive language. Alhaj et al. [4], proposes a novel text classification technique using improved particle swarm optimization for Arabic language. The authors evaluate the performance of their model using various metrics and compare it with other state-of-the-art models. The results show that their model outperforms other models in terms of accuracy and F1-score. Sultan et al. [21] evaluates the performance of shallow and deep learning algorithms for cyberbullying detection. The authors use three deep and six shallow learning algorithms and evaluate their performance in terms of accuracy and recall. The results show that bidirectional LSTM is the most efficient method for cyberbullying detection. Pawar et al. [15] emphasizes the importance of explainability in the decision-making process of deep learning models. The authors propose using interpretable features such as sentiment analysis, part-of-speech (POS) tagging [23], and topic modeling to extract linguistic and semantic patterns indicative of abusive language. They also propose using attention mechanisms and local interpretable model-agnostic explanations (LIME) to provide insights into the decision-making process of deep learning models. Pawar et al. [15] proposes a machine learning model for cyberbullying detection on Twitter. The authors use LIME to evaluate the performance of their model and provide explainability.

The papers discussed in this text focus on the detection of hate speech and cyberbullying using machine learning and natural language processing techniques. Bunde [6] proposes an artifact that integrates humans in the process of detecting and evaluating hate speech using explainable artificial intelligence (XAI). Cai et al. [7] propose an automatic misuse detector (MiD) for detecting potential bias in text classifiers and an end-to-end debiasing framework for text classifiers. Dewani et al. [8] propose a cyberbullying detection approach for analyzing textual data in the Roman Urdu language based on advanced preprocessing methods, voting-based ensemble techniques, and machine learning algorithms. Dewani et al. [9] perform extensive preprocessing on Roman Urdu microtext and analyze and uncover cyberbullying textual patterns in Roman Urdu using RNN-LSTM, RNN-BiLSTM, and CNN models. Herm et al. [10] conduct two user experiments to measure the tradeoff between model performance and explainability for five common classes of machine learning algorithms and address the problem of end user perceptions of explainable artificial intelligence augmentations.

The papers discussed the use of Explainable AI (XAI) methods to interpret deep learning models in Arabic Sentiment Analysis (ASA) and hate speech detection on social media platforms. Abdelwahab et al. [1] used Local Interpretable Model-agnostic Explanation (LIME) to demonstrate how the LSTM leads to the prediction of sentiment polarity within ASA. Ahmed and Lin [2] proposed a method for instance selection based on attention network visualization to detect hate speech. The approach uses active learning cycles to train the task using the result-label pairs and improves the model's accuracy. Babaeianjelodar et al. [5] built an explainable and interpretable high-performance model based on the XGBoost algorithm, trained on Twitter data, to detect hate and offensive speech. The paper uses Shapley Additive Explanations (SHAP) on the XGBoost models' outputs to make it explainable and interpretable compared to black-box models. These papers demonstrate the importance of XAI methods in interpreting deep learning models and improving the accuracy of hate speech and offensive speech detection. Ibrahim et al. [11] discuss the importance of explainability in hate speech detection models and propose a combination of XGBoost and logical LIME explanations for more logical results. Kouvela et al. [12] propose an explainable bot-detection approach for Twitter, which offers interpretable, responsible, and AI-driven bot identification. Mehta and Passi [13] demonstrate the potential of explainable AI in hate speech detection using deep learning models and the ERASER benchmark. Finally, Montiel-Vázquez et al. [14] provide a comprehensive study on the nature of empathy and a method for detecting it in textual communication, using a pattern-based classification algorithm for predicting empathy levels in conversations. These papers highlight the importance of not only developing accurate AI and ML models for detecting problematic content on social media platforms but also ensuring that these models are explainable and interpretable to maintain users' trust and understanding.

Pérez-Landa et al. [16] propose an XAI model for detecting xenophobic tweets on Twitter, which provides a set of contrast patterns describing xenophobic tweets to help decision-makers prevent acts of violence caused by such posts. Raman et al. [17] compare different hate and aggression detection algorithms, including machine learning models and deep learning models, and find that CNN+GRU static + Word2Vec embedding outperforms all other techniques. Sabry et al. [18] investigate the performance of a state-of-the-art architecture T5 and compare it with three other previous state-of-the-art architectures across five different tasks from two diverse datasets. They achieve near-state-of-the-art results on a couple of the tasks and use explainable artificial intelligence (XAI) to earn the trust of users. Shakil and Alam [20] propose a CNN-LSTM and NLP fusion strategy for characterizing malicious and non-malicious remarks with a word embedding technique and interpret the algorithms with an XAI-SHAP. These papers highlight the importance of developing accurate and interpretable AI and ML models for detecting problematic content on social media platforms. Shakil and Alam [19] propose a CNN-NLP fusion strategy for characterizing malicious and non-malicious remarks with a word embedding technique and interpret the algorithms with an XAI-SHAP. Their proposed architecture achieves a malicious comment classification accuracy of 99.75%, which is higher than previous work. Sultan et al. [21] evaluate shallow machine learning and deep learning methods for cyberbullying detection and find that bidirectional long-short-term memory is the most efficient method in terms of accuracy and recall. Wich et al. [25] develop an abusive language detection model leveraging user and network data to improve classification performance and integrate the explainable AI framework SHAP to assess the model's vulnerability toward bias and systematic discrimination reliably.

The papers provide insights into various XAI techniques and frameworks, including interpretable feature engineering, explainable deep learning models, and hybrid approaches, and provide case studies and ethical and privacy considerations. The papers also discuss XAI-driven cyberbullying intervention strategies, including real-time alert systems, explainable recommendations for moderators, and personalized feedback for users. Finally, the papers outline future directions, including technological advances, integration with other emerging technologies, addressing bias and fairness, and legal and regulatory implications.

2.1 Role of XAI in Cyberbullying Detection

The role of Explainable Artificial Intelligence (XAI) in cyberbullying detection is to address the limitations of traditional AI and ML approaches, particularly in terms of interpretability, trustworthiness, and accountability. XAI-driven models aim to provide clear, human-readable explanations for their decisions, which can help stakeholders better understand and trust these AI-driven systems.

In the context of cyberbullying detection, XAI can improve the interpretability of AI models by highlighting the specific features and reasoning behind a model's prediction. This increased transparency can help platform moderators, end-users, and policymakers gain insights into the underlying decision-making processes, enabling them to make more informed decisions about the actions they should take in response to detected instances of cyberbullying. For example, an explainable model may reveal that certain linguistic patterns or combinations of words are indicative of cyberbullying, helping stakeholders understand why a particular message was flagged.

Moreover, XAI-driven models can enhance trust in AI systems by providing evidence-based explanations for their predictions. This is particularly important when dealing with sensitive issues like cyberbullying, where the consequences of false positives or negatives can have significant impacts on the well-being of individuals. By offering explanations, XAI models enable stakeholders to verify and validate the system's predictions, ensuring that the model is accurately identifying instances of cyberbullying and not simply flagging content based on biases or other unrelated factors. XAI can facilitate accountability in AI-driven cyberbullying detection systems. When AI models provide clear explanations for their decisions, it becomes easier to identify and address any potential biases or errors in the model's predictions. This increased accountability can help build confidence among stakeholders, assuring them that the AI system is being used responsibly and ethically to combat cyberbullying.

2.2 XAI Models for Text Analysis

XAI models for text analysis aim to enhance the interpretability and transparency of AI-driven text classification systems, such as used in cyberbullying detection. These models combine AI techniques with explainability, enabling stakeholders to better understand the reasoning behind their predictions. Several XAI models for text analysis have emerged, including interpretable feature engineering, explainable deep learning models, and hybrid models.

Interpretable feature engineering focuses on creating meaningful and human-readable features that can be easily understood by users. These features often capture linguistic and semantic patterns indicative of the target phenomenon, such as cyberbullying. Techniques like sentiment analysis [22], part-of-speech tagging, and topic modeling can be used to extract interpretable features from text. By using interpretable features as input for machine learning classifiers, the model's decision-making process becomes more transparent, as stakeholders can directly examine the relationship between the features and the model's predictions.

Explainable deep learning models aim to address the black-box nature of traditional deep learning techniques, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), by providing insights into their decision-making processes. One approach is to use attention mechanisms, which allow the model to assign importance weights to different input elements, making it possible to visualize and understand which parts of the text contribute most to the model's prediction. Another approach involves local interpretable model-agnostic explanations (LIME) or layer-wise relevance propagation (LRP), which provide post-hoc explanations for individual predictions by approximating the deep learning model with a simpler, more interpretable model for a specific input.

Hybrid models combine multiple AI techniques, such as interpretable feature engineering, deep learning, and explainable AI methods, to create more powerful and transparent text classification systems. For example, a hybrid model may use deep learning to automatically extract complex patterns from text and interpretable feature engineering to create human-readable features. The model can then combine these features using an explainable classifier, such as an interpretable decision tree or rule-based system, which provides clear explanations for its predictions. This combination of techniques can lead to more effective and transparent models for text analysis in the context of cyberbullying detection.

2.3 Interpretability Metrics

Evaluating the effectiveness of XAI methods applied to cyberbullying or hate speech detection requires a combination of performance metrics and interpretability metrics. While performance metrics focus on the accuracy and generalizability of the AI models, interpretability metrics assess the quality of the explanations provided by the XAI methods. Some of the key interpretability metrics used to evaluate XAI methods for cyberbullying or hate speech detection include:

Fidelity measures the extent to which an explanation reflects the actual behavior of the AI model. High fidelity implies that the explanation accurately captures the model's decision-making process. It can be quantified by comparing the predictions made by the original model and the explanation method, using metrics such as R-squared, correlation coefficients, or mean squared error.

Consistency evaluates the degree to which explanations for similar instances are alike. A high consistency score indicates that the XAI method provides stable and coherent explanations across different instances. Consistency can be measured using clustering techniques, similarity measures, or by comparing the explanation output against a known ground truth.

Simplicity measures the complexity of the explanations provided by the XAI method. Ideally, explanations should be simple and easy for humans to understand. Simplicity can be quantified using metrics such as the number of features or rules in the explanation, the length of the explanation, or by evaluating the cognitive load required to comprehend the explanation.

Coverage assesses the proportion of instances for which the XAI method can provide meaningful explanations. A high coverage score indicates that the XAI method is capable of explaining a wide range of instances. Coverage can be computed as the percentage of instances for which the explanation method generates valid, non-trivial explanations.

Local faithfulness evaluates how well the explanation reflects the model's behavior for a specific instance within a local neighborhood. High local faithfulness implies that the explanation accurately captures the model's decision-making process for the given instance and its neighbors. It can be quantified by analyzing the changes in the model's output and explanation as small perturbations are introduced to the input.

Human evaluation involves subjective assessments of the explanations provided by the XAI method by domain experts, end-users, or other stakeholders. This evaluation can be conducted using surveys, interviews, or user studies, where participants rate the quality of the explanations based on factors such as understandability, usefulness, and trustworthiness.

By considering a combination of fidelity, consistency, simplicity, coverage, local faithfulness, and human evaluation, researchers can gain insights into the quality of the explanations provided by XAI methods and identify areas for improvement, ultimately contributing to the development of more transparent, trustworthy, and effective AI-driven solutions for detecting and addressing cyberbullying and hate speech.

2.4 Datasets for Cyberbullying Detection

Datasets for cyberbullying detection play a vital role in training, validating, and evaluating XAI methods in the context of text analysis. These datasets typically consist of annotated text samples from various online platforms, such as social media, forums, and messaging apps, labeled as cyberbullying or non-cyberbullying instances. It is essential for these datasets to be diverse, representative, and balanced to ensure the effectiveness and generalizability of the developed XAI models. Some of the widely used datasets for cyberbullying detection include:

Formspring.me dataset consists of over 12,000 user-generated questions and answers from the now-defunct social Q&A platform Formspring.me. The dataset contains binary labels for each post, indicating whether it is considered cyberbullying or not. The annotations were provided by multiple independent annotators, and their agreement was used to determine the final labels.

Twitter datasets have been created using Twitter data, where tweets are collected and annotated for cyberbullying or aggressive behavior. These datasets may contain various types of annotations, such as binary labels (e.g., bullying vs. non-bullying) or multi-class labels (e.g., offensive language, hate speech, or neutral). Twitter datasets often require extensive preprocessing and cleaning, as they may include noise, slang, abbreviations, and other challenges associated with social media text.

MySpace dataset is derived from the social networking site MySpace and contains over 80,000 comments from public profiles. The dataset is labeled using a binary classification of cyberbullying or non-cyberbullying instances, with annotations provided by human annotators.

Wikipedia Talk Page dataset consists of user comments from the talk pages of Wikipedia articles, where users discuss edits and other topics related to the articles. The dataset is annotated with multiple categories, such as personal attacks, harassment, or other aggressive behaviors, making it suitable for multi-class cyberbullying detection tasks.

ASKfm dataset is extracted from the social Q&A platform ASKfm and contains a collection of anonymous questions and answers. The dataset is labeled for binary cyberbullying detection, with annotations provided by human annotators.

3 XAI-Driven Cyberbullying Intervention Strategies

XAI-driven cyberbullying intervention strategies leverage the explainable artificial intelligence (XAI) methods to not only detect cyberbullying instances but also provide meaningful insights into the detected content, enabling stakeholders to develop targeted and effective intervention strategies. The following are some potential XAI-driven cyberbullying intervention strategies:

XAI methods can be used to generate personalized feedback to offenders, explaining why their content was flagged as cyberbullying or hate speech. By providing clear and understandable explanations, the offenders might gain a better understanding of the consequences of their actions and be encouraged to reconsider their behavior in future online interactions.

XAI-driven models can help empower bystanders by providing them with explanations regarding the detection of cyberbullying instances. Armed with this information, bystanders may feel more confident in intervening, either by reporting the abusive content, offering support to the victim, or directly addressing the offender in a constructive manner.

Explainable AI models can support human moderators and platform administrators in decision-making by providing insights into the reasons behind flagged content. These explanations can help them make more informed decisions on actions such as content removal, issuing warnings, or banning users, thus ensuring a safer and more inclusive online environment.

By analyzing the explanations provided by XAI methods, stakeholders can identify patterns and trends in cyberbullying behavior. These insights can be used to develop tailored preventive measures and educational resources, such as awareness campaigns, workshops, or online courses, that address the specific factors contributing to cyberbullying in a particular community or platform.

Explanations generated by XAI methods can be valuable for policymakers and legislators as they help to identify common patterns and trends in cyberbullying behavior. This information can be used to develop targeted policies and regulations that address the root causes of cyberbullying, ensuring more effective prevention and intervention strategies at a societal level.

XAI-driven cyberbullying intervention strategies can enhance the effectiveness of efforts to combat cyberbullying by providing clear and understandable explanations for AI-driven predictions. These explanations can inform targeted intervention strategies, such as personalized feedback, empowering bystanders, supporting moderators, tailored preventive measures, and informing policy and legislation, ultimately contributing to a safer and more inclusive online environment.

3.1 Real-Time Alert Systems

XAI-supported real-time cyberbullying and hate speech alert systems leverage explainable artificial intelligence (XAI) methods to provide real-time detection and explanations of cyberbullying or hate speech instances on various online platforms. These systems aim to improve the transparency, trustworthiness, and effectiveness of AI-driven content moderation and intervention strategies. Below, we discuss key aspects of XAI-supported real-time cyberbullying and hate speech alert systems: By utilizing advanced AI models, such as deep learning and natural language processing techniques, these alert systems can process and analyze large volumes of text data from various online platforms in real-time. This enables the rapid identification of potential instances of cyberbullying or hate speech and allows for timely interventions to minimize harm to the affected individuals.

XAI-supported alert systems incorporate explainable AI models, such as interpretable feature engineering, explainable deep learning models, or hybrid models, which provide understandable explanations for their decisions. These explanations can offer insights into the reasoning behind the model's predictions and enhance the transparency and trustworthiness of the alert system. When a potential instance of cyberbullying or hate speech is detected, the XAI-supported alert system can generate real-time notifications to relevant stakeholders, such as platform administrators, moderators, or even the affected individuals. These alerts can be accompanied by explanations generated by the XAI models, providing stakeholders with valuable context to inform their intervention strategies. By understanding the explanations provided by the XAI models, stakeholders can develop customized intervention strategies to address the detected instances of cyberbullying or hate speech. These strategies can range from automated actions, such as content filtering or temporary content suspension, to more nuanced human intervention, such as contacting the involved parties, offering support to the victims, or educating the offenders.

XAI-supported real-time alert systems can facilitate continuous improvement and adaptability by allowing stakeholders to evaluate the effectiveness of the AI models and their explanations. By analyzing the generated explanations, stakeholders can identify potential areas for improvement in the models or the intervention strategies, ensuring the system remains effective and relevant as new forms of cyberbullying or hate speech emerge. XAI-supported real-time cyberbullying and hate speech alert systems offer a promising approach to enhancing the transparency, trustworthiness, and effectiveness of AI-driven content moderation and intervention strategies. By providing real-time detection and explanations for potential instances of cyberbullying or hate speech, these systems enable stakeholders to develop targeted and timely intervention strategies, ultimately contributing to a safer and more inclusive online environment.

An example XAI-supported real-time cyberbullying and hate speech alert system operation scenario is given in Fig. 1.

This sequence diagram includes five participants: the User (U), the AI Model (A), the XAI Method (X), the Alert System (AS), and the Moderator (M). The process starts when a user posts content on a platform. The AI model analyzes the posted content, and if the content is flagged as potentially harmful, the XAI method is employed to generate an explanation for the AI model's decision. The flagged content and the generated explanation are then passed to the alert system, which sends a real-time alert along with the explanation

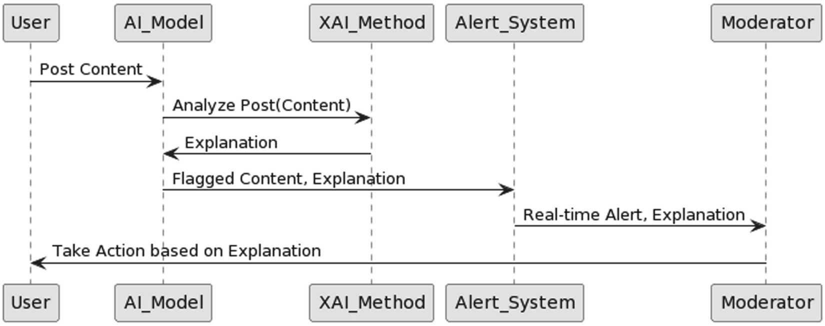

Fig. 1. An example XAI-supported real-time cyberbullying and hate speech alert system operation scenario.

to the moderator. The moderator then takes appropriate action based on the explanation provided.

3.2 Explainable Recommendations for Moderators

Explainable Artificial Intelligence (XAI) can be employed to provide explainable recommendations for social media platform moderators, enhancing the transparency, trustworthiness, and effectiveness of content moderation processes. By combining advanced AI models with interpretable explanations, XAI-supported recommendations can help moderators better understand the reasoning behind the suggested actions, allowing them to make more informed decisions in managing online content. Below, we discuss how XAI could be employed to provide explainable recommendations for social media platform moderators:

By extracting and highlighting interpretable features from the text data, such as keywords, phrases, or sentiment scores, XAI can provide moderators with meaningful insights into the factors contributing to the AI model's predictions. This enables moderators to understand the context of the flagged content and make more informed decisions about the appropriate actions to take.

XAI methods, such as LIME or SHAP, can generate locally faithful explanations for individual instances of flagged content. These explanations can help moderators understand the specific factors that led the AI model to classify a particular piece of content as cyberbullying, hate speech, or otherwise inappropriate. This allows moderators to assess the relevance and accuracy of the model's predictions and make informed decisions based on the provided explanations.

Attention mechanisms in deep learning models can be used to generate explanations by highlighting the most relevant parts of the input data (e.g., words or phrases) that contribute to the model's predictions. By visualizing the attention weights, moderators can gain insights into the decision-making process of the AI model and understand which aspects of the content were deemed problematic.

XAI can provide contextual and temporal explanations by considering the broader context of the flagged content, such as user profiles, interaction histories, or the timing of the posts. This additional information can help moderators understand the larger context in which the content was posted, allowing them to make more informed decisions about the appropriate intervention strategies.

XAI can be employed to generate explainable recommendations for various intervention strategies, such as content removal, user warnings, or account suspensions. By providing insights into the factors contributing to the AI model's predictions, XAI can help moderators understand the potential risks and benefits associated with different intervention strategies, enabling them to make more informed decisions that balance the need for a safe online environment with the preservation of freedom of expression.

Employing XAI to provide explainable recommendations for social media platform moderators can enhance the transparency, trustworthiness, and effectiveness of content moderation processes. By offering interpretable insights into the AI model's decision-making process, XAI can empower moderators to make more informed decisions in managing online content, ultimately contributing to a safer and more inclusive online environment.

A sequence diagram representing an example Explainable Recommendation generation for Moderators scenario is presented in Fig. 2.

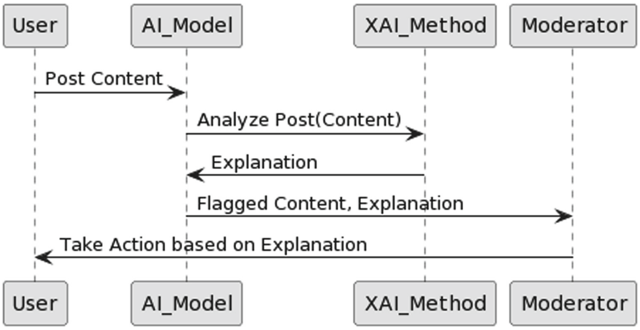

Fig. 2. Explainable Recommendation generation for Moderators scenario.

This sequence diagram includes four participants: the User (U), the AI Model (A), the XAI Method (X), and the Moderator (M). The process starts when a user posts content on a platform. The AI model then analyzes the posted content, and if the content is flagged as potentially harmful, the XAI method is employed to generate an explanation for the AI model's decision. The flagged content and the generated explanation are then passed to the moderator, who takes appropriate action based on the explanation provided.

4 Conclusion

Application of Explainable Artificial Intelligence (XAI) methods in combating cyberbullying and hate speech holds significant promise for enhancing the effectiveness, transparency, and trustworthiness of AI-driven solutions. This paper has provided an overview of traditional approaches to cyberbullying detection, the role of XAI in this context, XAI models for text analysis, datasets used for cyberbullying detection, interpretability metrics, and real-life examples of cyberbullying detection using XAI methods.

As cyberbullying and hate speech continue to pose significant challenges for individuals and communities worldwide, the integration of XAI methods in prevention and mitigation efforts can contribute to a safer and more inclusive online environment. By enhancing the transparency and interpretability of AI-driven solutions, XAI can empower stakeholders to make more informed decisions and develop more effective, targeted, and responsible strategies for addressing harmful online behaviors.

Future research should focus on the development of novel XAI techniques, the integration of multimodal data, improving contextual understanding, real-time explanations, personalized and targeted interventions, and collaboration with human experts. Additionally, addressing ethical and legal considerations is vital to ensure the responsible and equitable application of XAI methods in combating cyberbullying and hate speech.

References

1. Abdelwahab, Y., Kholief, M., Sedky, A.A.H.: Justifying Arabic text sentiment analysis using explainable AI (XAI): lasik surgeries case study. Information **13**(11), 536 (2022)
2. Ahmed, U., Lin, J.C.: Deep explainable hate speech active learning on social-media data. IEEE Trans. Comput. Soc. Syst. (2022)
3. Aldjanabi, W., Dahou, A., Al-Qaness, M.A.A., Elaziz, M.A., Helmi, A.M., Damaševičius, R.: Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. Informatics **8**(4), 69 (2021)
4. Alhaj, Y.A., et al.: A novel text classification technique using improved particle swarm optimization: a case study of Arabic language. Future Internet **14**(7), 194 (2022)
5. Babaeianjelodar, M., et al.: Interpretable and high-performance hate and offensive speech detection. In: Chen, J.Y.C., Fragomeni, G., Degen, H., Ntoa, S. (eds.) HCII 2022. LNCS, vol. 13518, pp. 233–244. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-21707-4_18
6. Bunde, E.: AI-assisted and explainable hate speech detection for social media moderators - a design science approach. In: Annual Hawaii International Conference on System Sciences, vol. 2020-January, pp. 1264–1273 (2021)
7. Cai, Y., Zimek, A., Wunder, G., Ntoutsi, E.: Power of explanations: towards automatic debiasing in hate speech detection. In: 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA 2022) (2022)
8. Dewani, A., Memon, M.A., Bhatti, S.: Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data. J. Big Data **8**(1), 160 (2021). https://doi.org/10.1186/s40537-021-00550-7
9. Dewani, A., et al.: Detection of cyberbullying patterns in low resource colloquial roman urdu microtext using natural language processing, machine learning, and ensemble techniques. Appl. Sci. **13**(4), 2062 (2023)

10. Herm, L., Heinrich, K., Wanner, J., Janiesch, C.: Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. Int. J. Inf. Manag. **69**, 10253 (2023)
11. Ibrahim, M.A., et al.: An explainable AI model for hate speech detection on Indonesian twitter. CommIT J. **16**(2), 175–182 (2022)
12. Kouvela, M., Dimitriadis, I., Vakali, A.: Bot-detective: an explainable twitter bot detection service with crowdsourcing functionalities. In: 12th International Conference on Management of Digital EcoSystems (MEDES 2020), pp. 55–63 (2020)
13. Mehta, H., Passi, K.: Social media hate speech detection using explainable artificial intelligence (XAI). Algorithms **15**(8), 291 (2022)
14. Montiel-Vázquez, E.C., Ramírez Uresti, J.A., Loyola-González, O.: An explainable artificial intelligence approach for detecting empathy in textual communication. Appl. Sci. **12**(19), 9407 (2022)
15. Pawar, V., Jose, D.V., Patil, A.: Explainable AI method for cyber bullying detection. In: 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC 2022) (2022)
16. Pérez-Landa, G.I., Loyola-González, O., Medina-Pérez, M.A.: An explainable artificial intelligence model for detecting xenophobic tweets. Appl. Sci. **11**(22), 10801 (2021)
17. Raman, S., Gupta, V., Nagrath, P., Santosh, K.C.: Hate and aggression analysis in NLP with explainable AI. Int. J. Pattern Recognit. Artif. Intell. **36**(15), 2259036 (2022)
18. Sabry, S.S., Adewumi, T., Abid, N., Kovacs, G., Liwicki, F., Liwicki, M.: Hat5: hate language identification using text-to-text transfer transformer. In: International Joint Conference on Neural Networks, vol. 2022-July (2022)
19. Shakil, M.H., Alam, M.G.R.: Hate speech classification implementing NLP and CNN with machine learning algorithm through interpretable explainable AI. In: 2022 IEEE Region 10 Symposium (TENSYMP 2022) (2022)
20. Shakil, M.H., Rabiul Alam, M.G.: Toxic voice classification implementing CNN-LSTM & employing supervised machine learning algorithms through explainable AI-Shap. In: 4th IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET 2022) (2022)
21. Sultan, D., et al.: Cyberbullying-related hate speech detection using shallow-to-deep learning. Comput. Mater. Cont. **74**(1), 2115–2131 (2023)
22. Tesfagergish, S.G., Damaševičius, R., Kapočiūtė-Dzikienė, J.: Deep Learning-Based Sentiment Classification of Social Network Texts in Amharic Language, Communications in Computer and Information Science, vol. 1740. CCIS (2022)
23. Tesfagergish, S.G., Kapočiūtė-Dzikienė, J.: Part-of-speech tagging via deep neural networks for northern-ethiopic languages. Inf. Technol. Control **49**(4), 482–494 (2020)
24. Venckauskas, A., Karpavicius, A., Damasevicius, R., Marcinkevicius, R., Kapociute-Dzikiene, J., Napoli, C.: Open class authorship attribution of lithuanian internet comments using one-class classifier. In: 2017 Federated Conference on Computer Science and Information Systems (FedCSIS 2017), pp. 373–382 (2017)
25. Wich, M., Mosca, E., Gorniak, A., Hingerl, J., Groh, G.: Explainable abusive language classification leveraging user and network data. In: Dong, Y., Kourtellis, N., Hammer, B., Lozano, J.A. (eds.) Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track. ECML PKDD 2021. LNCS, vol. 12979, pp. 481–496. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86517-7_30