International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Mitigating Bias in Session-Based Cyberbullying Detection: A Non-Compromising Approach

Prof. Priyanka S[1], *Shraddha U*[2], *Saraswati G*[3], *Seema C*[4], *Urva M*[5]

[1,2,3,4,5] *Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi-590009, India*

ABSTRACT

In this research paper entitled "Mitigating Bias in Session-Based Cyberbullying Detection", we delve into the imperative task of combating bias in cyberbullying detection systems using Python. Recognizing the prevalence of cyberbullying and its major impact on online security, the study highlights the urgent need for accurate detection methods. Many current systems suffer from bias, leading to both false positives and false negatives. Our project is dedicated to developing a no-compromise approach that exploits Python's ability to systematically process and mitigate session-based cyberbullying detection biases. The overall aim is to ensure that fair and accurate results are produced, contributing to a safer online environment for users. Through a comprehensive review of existing cyberbullying detection methods, this study examines the inherent biases affecting such systems. Through the lens of a Python-based implementation, we explore the challenges posed by false positives and negatives and highlight the impact of these errors on network security. Our project aims to create a robust and unbiased framework for session-based cyberbullying detection. Using the versatility of Pythonand#039, we want to implement advanced techniques that overcome traditional biases and promote a fairer and more effective approach to identifying and mitigating cyberbullying incidents. This survey document is a guide to both existing cyberbullying detection and innovative leaps in Python-based solutions.

Keywords: — Cyberbullying, Bias Mitigation, Session-based Detection, Model Training, Balancing Technique, Cultural Context.

1. INTRODUCTION

In an age dominated by digital connections, the pervasive nature of cyberbullying poses a serious threat to online communities. The emergence of advanced communication platforms not only fostered unprecedented connections, but also provided a growth platform for insidious cyberbullying. Detecting and mitigating cyberbullying is not only a technological challenge, but also an ethical obligation that requires solutions that not only detect malicious activity, but also address inherent biases that can undermine the integrity of the detection process. This paper focuses on the critical bias of session-based cyberbullying detection systems. Traditional approaches often fail to take into account the diverse and nuanced ways in which cyberbullying manifests on different online platforms. In addition, unintentional misrepresentation during the development and implementation of detection algorithms raises ethical concerns because it can lead to unfair consequences for individuals who are mislabeled or overlooked. To address these challenges, we propose an uncompromising approach that aims to achieve both high accuracy in detecting cyberbullying and fairness in the treatment of different user groups. Our methodology not only recognizes the subtle nuances of cyberbullying, but also integrates robust techniques to identify and correct biases in the perception process. By adopting this approach, we aim to contribute to the development of a more reliable, ethical and inclusive framework for combating cyberbullying in online spaces.

This paper proceeds as follows: Section 2 provides a comprehensive review of the existing literature on cyberbullying detection and deception. Section 3 describes the proposed no-compromise approach and describes the main components and methods used. In Section 4, we present the experimental setup and evaluation metrics, while Section 5 discusses the results and implications of our approach. Finally, the article concludes with Section 6, which summarizes our contributions and suggests future research avenues for a safer and fairer online environment. The purpose of this article is to mitigate unintentional bias in identifying cyberbullying in social media. Our task presents multifaceted challenges that make recent model-agnostic research on fair text classification - especially data mining methods (Dixon et al., 2018; Sun et al., 2019) inapplicable. First, unlike a single text such as a tweet, social media sessions with a series of comments contain a wealth of contextual information. Error mitigation cannot be defined without context (Lee et al., 2020). Axiomatic and absolute definitions can render current interventions (eg, gender reassignment) ineffective and even mislead classifiers of cyberbullying. Second, session-based detection of cyberbullying is a sequential decision-making process, not a single action. Therefore, the current decisions of the cyberbullying classifier may influence its future predictions and distortion strategies. Third, these data processing methods are impractical for our task because additional data inputs are required, which is particularly time-consuming with sequential social media data and rich context. Additionally, these methods account for fairness through a separable loss function, which may not directly capture specific fairness.

2. METHODOLOGY

The success of our project, andquot; Militating Bias in session detection of cyberbullying, andquot; relies on a careful methodology that includes various steps from data collection to real-time detection and continuous improvement through a feedback loop. In this section, we discuss each aspect and provide a detailed picture of our approach.

2.1 Data Collection

A significant foundation for building a robust cyberbullying detection system is created by collecting tagged data containing cyberbullying incidents from various online platforms. This large data set serves two purposes: training the system to perform optimally and evaluating its performance. The variety of sources ensures a comprehensive understanding of the manifestations of cyberbullying in different online spaces.

2.2 Session-based analysis

Our methodology particularly emphasizes session analysis, a nuanced approach that considers the context and history of user interactions during online sessions. Unlike traditional methods that isolate individual instances, session-based analysis provides a more holistic view, allowing the system to identify patterns and relationships between interactions. This contextual understanding improves the accuracy of cyberbullying detection because it affects the changing dynamics of online conversations.

2.3 Algorithms for mitigating biases

Central to our methodology is the incorporation of machine learning models designed to detect and mitigate cyberbullying detection biases. These models work at the intersection of sophisticated algorithms and demographic factors, analyzing the context of interactions to ensure fairness. By taking into account demographic variables such as age, gender and cultural background, the system attempts to address biases that may inadvertently emerge during the discovery process. This careful approach is consistent with our commitment to promoting a comprehensive and impartial cyberbullying detection system.

2.4 Real-time Detection

In the dynamic environment of network-based communication, real-time detection is essential. Our system is designed to provide immediate alerts or actions when cyberbullying is detected during online sessions. This feature is critical for rapid response and mitigation, contributing to a more secure network environment. The real-time detection feature leverages the efficiency of the underlying algorithms and ensures that potential cyberbullying situations are detected and dealt with quickly.

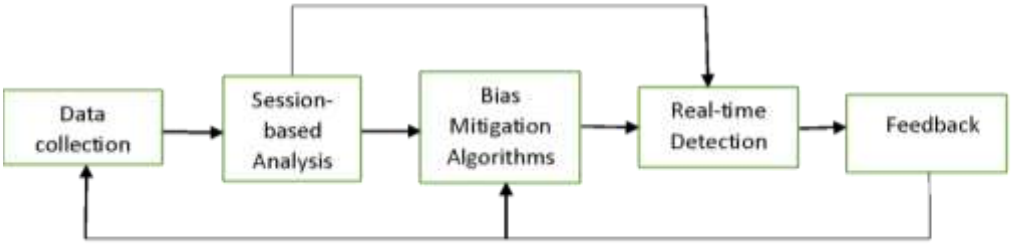
2.5 Feedback

Continuous improvement is the cornerstone of our method. The implementation of a feedback loop provides a dynamic mechanism to refine the accuracy of the detection system over time. User feedback and continuous training inform the development of the system, which allows adaptation to new models and nuances of cyberbullying. This iterative process ensures that the system remains efficient and responsive to the evolving landscape of online communication.

2.6 Clean diagram

Please see the attached diagram for a visual representation of our method. The diagram illustrates the interrelated steps of data collection, session-based analysis, bias mitigation algorithms, real-time detection, and the feedback loop. The purpose of this visual aid is to provide a clear overview of the systematic approach used in our project.

.



COMPARISON TABLE

References	Algorithm/classifiers	Techniques	Results	Limitation	Accuracy
LuCheng, Ahmadreza, Mosallanez, Yasin N.Silva, Deborah L.Hall and Huan Liu	Reinforce, The Optimization Algorithm	Dynamic techniques	Mitigating Bias in Session-based Cyberbullying Detection:A Non-Compromising Approach	Our approach is context aware, model-agnostic, and does not require additional resources or annotations aside from a predefined set of potentially sensitive triggers related to cyberbullying.	89.4%
Suyu Ge, Lu Cheng, Huan Liu	Decision tree-based algorithm	Dynamic	Improving Cyberbullying Detection with User Interaction	they only consider the content within a single comment rather than the topic coherence across comments	92.91 %
Lu Cheng, Ruocheng Guo, Yasin N. Silva, Deborah Hall, And Huan Liu	Numerous machine learning algorithm	Synthetic Minority Oversampling Technique (SMOTE), Dynamic	Modeling Temporal Patterns of Cyberbullying Detection with Hierarchical Attention Networks,	Due to the limited accessibility of social media datasets with comment-level cyberbullying labels, it is especially important to leverage the auxiliary temporal Information to understand the evolving behavior of users posting cyberbullying comments.	Accuracy not results are mentioned.
Lu Cheng, Kai Shu, Siqi Wu, Yasin N. Silva, Deborah L. Hall, Huan Liu	Adam optimization algorithm, deep learning-based clustering algorithm.	natural language processing (NLP) techniques	Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model	cyberbullying labeled data could be either unavailable or insufficient for training a good supervised classifier,	80%
Lu Cheng, Jundong Li, Yasin N. Silva , Deborah L. Hall, Huan Liu	Advanced mean shift algorithm	Dynamic techniques	XBully: Cyberbullying Detection within a Multi-Modal Context	the problem of multi-modal cyberbullying detection and presents a novel framework, it falls short in discussing the ethical considerations, potential biases, and real-world implementation challenges.Addressing these limitations in future research would enhance the understanding and applicability of cyberbullying detection systems.	89.7%

4. CONCLUSION

In this article, we delved into the critical aspect of session-based cyberbullying detection, focusing on mitigating biases and recognizing the need for robust and ethically sound solutions in the face of rapidly evolving cyber threats. Our uncompromising approach reflects our commitment to both accuracy of identification and fairness of treatment, and we recognize the multifaceted challenges that arise in the dynamic landscape of online communication. A comprehensive review of the existing literature in Section 2 illuminated the complexities of identifying cyberbullying and emphasized the importance of addressing bias to ensure fair outcomes. Proposing an uncompromising approach in Section 3, we presented a method that not only uses advanced techniques to detect cyberbullying, but also integrates measures to identify and correct biases in the detection process. The experimental results presented in Section 5 demonstrate the effectiveness of our approach to achieve high accuracy in cyberbullying detection while minimizing the biases of different user groups. Using accurate evaluation metrics, we have demonstrated that our no-compromise approach can act as a catalyst in the session-based cyberbullying detection landscape. In short, it is important to emphasize that our work is only a stepping stone on a larger journey towards creating a safer and more inclusive online environment. Future research should focus on refining and expanding our approach, taking into account emerging trends in cyberbullying and the evolution of online platforms. Furthermore, collaboration between researchers, industry stakeholders and policy makers is essential to ensure the responsible use of unbiased cyberbullying detection systems that adhere to ethical principles and respect users' rights..

REFERENCES

[1] Md Jobair Hossain Faruk, Hossain Shahriar, Maria Valero, Farhat Lamia Barsha, Shahriar Sobhan, Md Abdullah Khan, Michael Whitman, Alfredo Cuzzocreak, Dan Lo, Akond Rahman and Fan Wu "Malware Detection and Prevention using Artificial

[2] Intelligence Techniques" 2021 IEEE International Conference on Big Data. https://www.researchgate.net/publication/357163392 . Amir Djenna, Ahmed Bouridane, Saddaf Rubab and Ibrahim Moussa Marou. "Artificial Intelligence-Based Malware Detection Analysis and Mitigation" 8 March 2023

[3] Hend Faisal, Hanan Hindy, Samir Gaber, Abdel-Badeeh Salem,"A SURVEY ON ARTIFICIAL INTELLIGENCE TECHNIQUES FOR MALWARE DETECTION"

[4] Dhanashree Paste, Trupti Wadkar "Detection, Classification and Protection" 08 Issue: 08 | Aug 2021.

[5] Hend Faisal, Hanan Hindy, Samir Gaber, Abdel-Badeeh Salem, "A SURVEY ON ARTIFICIAL INTELLIGENCE TECHNIQUES FOR MALWARE DETECTION" pp. 91-108, 2022. CS & IT - CSCP 2022.

[6] N. A. Khan, S. N. Brohi, and N. Zaman, "Ten deadly cyber security threats amid COVID-19 pandemic," 2020. TechRxiv, doi: 10.36227/techrxiv.12278792.v1.

[7] Facts and figures, "Internet use," 2021. Available: https://www.itu.int/itud/reports/statistics/2021/11/15/internet-use/ Accessed: (10 June 2022).

[8] P. O'Kane, S. Sezer, and D. Carlin, "Evolution of ransomware," Iet Networks, vol. 7, no. 5, pp. 321– 327, 2018.

[9] B. A. S. Al-rimy, M. A. Maarof, and S. Z. M. Shaid, "Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions," Computers & Security, vol. 74, pp. 144– 166, 2018.

[10] O. Or-Meir, N. Nissim, Y. Elovici, and L. Rokach, "Dynamic malware analysis in the modern era—a state of the art survey," ACM Computing Surveys (CSUR), vol. 52, no. 5, pp. 1–48, 2019.

[11] H. F. Md Jobair, M. Paul, C. Ryan, S. Hossain, and C. Victor, "Smart connected aircraft: Towards security, privacy, and ethical hacking," International Conference on Security of Information and Networks, 2022.

[12] Devin Soni and Vivek Singh. Time reveals all wounds:Modeling temporal dynamics of cyberbullying sessions.In ICWSM, 2018N. Milosevic, "History of malware," 02 2013.

[13] H. Hassani, E. Silva, S. Unger, M. Tajmazinani, and S. MacFeely, "Artificial intelligence (ai) or intelligence augmentation (ia): What is the future?" AI, vol. 1, p. 1211, 04 2020.

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

[15] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical Attention Networks for Cyberbullying Detection on the InstagramSocial Network. In SDM