VirtuGuard: Ethically Aligned Artificial Intelligence Framework for Cyberbullying Mitigation

Min Wang*†, Christine Boshuijzen-van Burken†, Nan Sun†, Shabnam Kasra Kermanshahi†, Yu Zhang†, and Jiankun Hu†
*University of Canberra, Australia
min.wang@canberra.edu.au
†University of New South Wales, Canberra, Australia
{c.vanburken, nan.sun, s.kasra_kermanshahi, m.yuzhang, jiankun.hu}@unsw.edu.au

***Abstract*—Cyberbullying has become a concerning issue in contemporary society with the widespread use of digital communication tools and social media platforms. The impacts of cyberbullying can be far-reaching, especially for certain groups such as children and teenagers. This work aims to mitigate cyberbullying in an ethically appropriate manner with careful consideration of transparency, explainability, privacy protection, contextual understanding, and continuous monitoring and improvement. We propose an ethically aligned artificial intelligence framework for cyberbullying detection and analysis. The framework provides four core functions: 1) detecting cyberbullying comments with explanations; 2) building and enriching an evolutionary cyberbullying knowledge map with detected instances and external ethics resources; 3) constructing and maintaining a cyberbullying instance network; and 4) performing analytics and recommendation (e.g., mental aid support) based on the knowledge map and instance network. The output knowledge map, instance network, and analysis report collectively offer useful insights for policymakers, regulators, ethicists, industry stakeholders and researchers, facilitating the establishment of global standards and fostering collaborative efforts in addressing cyberbullying. We keep human in the loop and also ensure that user privacy is well protected.**

***Index Terms*—cyberbullying, ethics, EU Act, AI, deep learning, data privacy**

I. Introduction

Cyberbullying has become a major problem in today's society, with devastating effects for teenagers [1], a group that spends almost half their awaken time on devices. This project aims to mitigate online bullying in an ethically appropriate manner, using machine learning techniques, with a direct link to data analytics, knowledge management and user privacy. There are deep learning tools for detecting cyberbullying, but we are proposing a system with ethics consideration and knowledge management, which is novel and requires cross-disciplinary expertise. Our research is timely with a view to the recent adoption of the Ethical AI EU Act [2].

The field of cyberbullying detection has witnessed substantial growth in recent years, driven by the urgent need to address its widespread and detrimental impacts, which transcend age, gender, and geographical boundaries [3]. Current research focuses on machine learning based detection [4], [5]. Spanning from 2011 to 2023, there has been a profound emphasis on leveraging machine learning techniques for the identification of cyberbullying [6], [7]. Underscoring a tech-centric strategy in tackling cyberbullying, highlighting a shift towards data-driven solutions. However, cyberbullying detection involves handling sensitive personal data and potentially impactful interventions. A human-machine teaming approach has the potential to ensure that these activities are conducted ethically, respecting privacy and minimizing harm. Motivated by this, we seek to create a system that not only detects cyberbullying but is also deeply rooted in understanding and mitigating the human impact of cyberbullying. Therefore, we propose VirtuGuard, which embeds cyberbullying detection, knowledge management, data analysis and recommendation modules into one system.

The advantage of VirtuGuard is that it is not only capable of detecting cyberbullying instances and providing explanatory power, but also provides analytical capabilities to build an evolutionary knowledge graph of cyberbullying and build a network of bullying instances, with ethical considerations. Detection predictions, knowledge graphs, and instance networks are used to generate advisory reports (e.g., for government regulation) and recommendations (e.g., recommending psychological assistance support information to users experiencing bullying).

II. VirtuGuard Conceptual Framework

VirtuGuard consists of five main modules, which are cyberbully detection AI, cyberbully knowledge management AI, cyberbully instance analytics AI, analytics and recommendation AI, and a privacy protection component (refer to Fig. 1).

A. Privacy Protection

To protect user privacy, we adopt anonymization and pseudonymization in the first step to ensure that it is difficult to trace back to individual users. Specifically, direct identifiable information (such as names, phone numbers, and email addresses) is removed from the data, and other sensitive text (such as usernames and locations) is replaced with tokens. We also introduce randomization to further obscure the data. This involves shuffling the order of records and replacing specific terms with randomly generated contents. The above procedures ensure that personally identifiable information is anonymized or pseudonymized during the detection and anal-

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00271

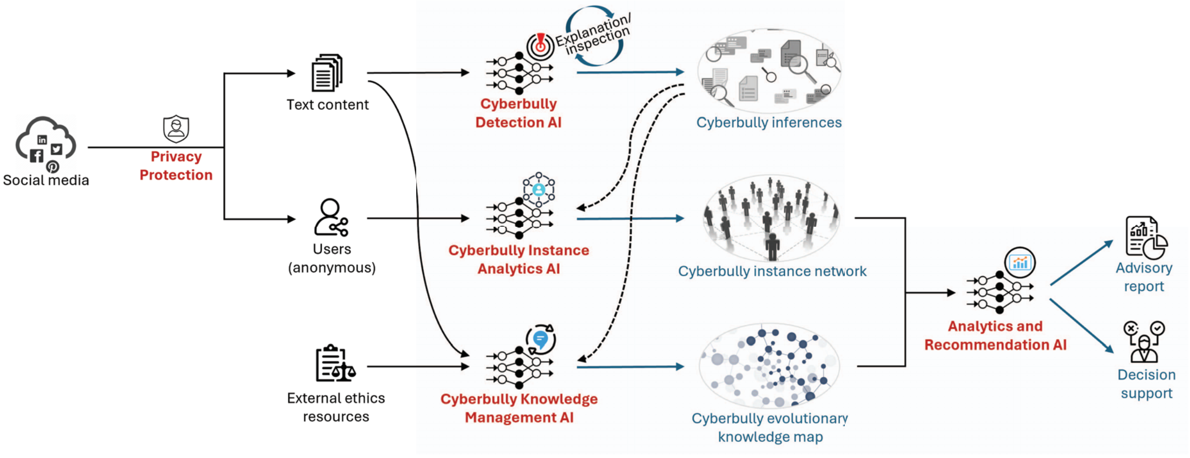
Social media
Privacy Protection
Text content
Users (anonymous)
External ethics resources
Explanation/ inspection
Cyberbully Detection AI
Cyberbully Instance Analytics AI
Cyberbully Knowledge Management AI
Cyberbully inferences
Cyberbully instance network
Cyberbully evolutionary knowledge map
Analytics and Recommendation AI
Advisory report
Decision support

Fig. 1. Schematic diagram of VirtuGuard, an ethically aligned artificial intelligence framework for cyberbullying mitigation.

ysis process. Hence, it prevents the direct association of data with individual users.

B. Cyberbullying Detection

A state-of-the-art ensemble model [3] is adopted for the cyberbullying detection AI module in our system. The ensemble model incorporates three second and third generation transformer models, which are DeBERTa (successor of BERT), GPT-Neo (lite version of GPT-2), and ALBERT (lite version of DistilBERT). DeBERTa, or Decoding-enhanced BERT with Disentangled Attention, is an advanced language model that improves upon the BERT architecture by incorporating disentangled attention mechanisms and enhanced decoding strategies, enhancing its capabilities in understanding context and relationships in natural language. GPT-Neo is a scaled-down version of OpenAI's GPT architecture, designed for efficient large-scale NLP tasks. ALBERT, or A Lite BERT, is an optimized and parameter-efficient variant of the BERT (Bidirectional Encoder Representations from Transformers) model, leveraging techniques such as shared layer parameterization and cross-layer parameter sharing to achieve impressive performance gains with fewer parameters. The three models were trained independently, then probability averaging was adopted for ensemble learning.

We use the IEEE Fine-Grained Cyberbullying Dataset (FGCD) [4] for training. This dataset offers cyberbullying cases distinguished by multiple traits, including age, ethnicity, gender, religion and other (bullying cases that do not match the specific traits), allowing multi-dimensional analysis. Hence, our ensemble model trained on the FGCD dataset provides trait identification while detecting cyberbullying. Meanwhile, the ensemble model provides class probabilities and inference that allow for human inspection, validation and interpretation.

C. Cyberbullying Instance Analytics

Relationships amongst the instances (anonymized users) are analyzed in this module based on their mutual interactions. These interactions can be reflected by information such as number of "likes" on the posts of a user, number of "followings" and "followers", and number of "sharings" and "forwardings". With the assistance of the cyberbullying inferences, the instance analytics AI identifies those influencing and raising cyberbullying instances based on the above interaction features, and meanwhile investigating how the cyberbullying instances have developed and have been influenced by other instances using the PageRank algorithm.

D. Cyberbullying Knowledge Management

This module aims to facilitate effective human-AI teaming by generating an evolutionary knowledge map of cyberbullying. The knowledge map integrates relevant inferences and text content retrieved from social media and external resources such as regulations (e.g., the EU AI Act) and mental aid information. Specifically, given the detected cyberbullying keywords, this module firstly clusters the keywords with early time stamps into topics based on semantic similarity, and then a hierarchical agglomerative clustering approach is used to create a knowledge map demonstrating the development of these cyberbullying topics across time. The nodes of the map refer to the topics; size of nodes denotes their occurrence frequency; and direction of edges represents their evolutionary relationship. The external ethics resources such as relative regulations or mental support services are then linked to the topics based on the context relevance. The implementation of this module draws on the K3S system [8].

E. Analytics and Recommendation

With the aim of providing recommendations to end users, this module takes the cyberbully evolutionary knowledge map and instance network as input for strategic analytics. A heterogeneous network is built which consists of three types of nodes including cyberbullying instances, topics and ethics resources which are cross-connected based on their relationships. A random walk-based approach [9] is used to identify the significant cyberbully topics considering the instances who are posting content around these topics, and thereby recommending relevant ethics resources to reduce the posts and help the victims.

III. Discussion

Interpretability. Deep learning models are often opaque in terms of model interpretability, inference, and the impact of large training sets on ethics and bias. Layer visualization, activation mapping, and sensitivity analysis can be employed to understand how the model derive the predictions [10]. These information provides explanations in different aspects to help validating the inference and facilitate human inspection. Moreover, the interpretability capacity is further strengthened by the knowledge management AI and cyberbullying instance analytics AI. These two AI modules build evolutionary knowledge maps and accumulate cyberbully instances based on the inference, which can be seen as 'explanations' of the inference from a knowledge extraction perspective. While it presents challenges, particularly in terms of complexity and the balance between interpretability and accuracy, it remains a key component in developing AI systems that are both effective and trustworthy.

Biases. When training the cyberbullying detection model, the impact of multiple biases should be carefully considered. This involve sampling bias (e.g., lack of randomness, or disposition towards certain groups), label bias (e.g., imbalanced labels), user engagement bias (e.g., over-representation of prominent users/perspectives), platform bias (e.g., bias towards the behaviors of certain community), and representation bias [11]. To address these biases, a carefully collected dataset is needed that takes into account balanced labels, stratified sampling, a broad range of users, diverse groups, and balanced representation of different demographics. Meanwhile, as the distribution of cyberbullying behaviors can change over time, update mechanisms are needed to adapt the model to evolving online behaviors and capture new patterns of cyberbullying.

Cultural/linguistic Domination. Our models are based on English language. Language embeds cultural values, including what counts as bullying. Hence our dataset may be biased towards bullying instances that are recognized as such in English language. Additionally, Wasseem found that datasets may be annotated differently depending on who annotates the data, for example amateurs classify items as bullying more than experts do [12].

Privacy-Preserving Model Training. User privacy is a serious concern in cyberbullying detection and mitigation, especially for sensitive groups such as children and teenagers. Currently, most of the models are trained on raw data from clients in a central server. To preserve privacy, differential privacy techniques such as adding Laplace noise or Gaussian noise during gradient computation or model updates can be implemented to help prevent the model from memorizing specific details about individual users. Moreover, secure multi-party computation and federated learning can be implemented to allow training a model collaboratively across decentralized devices without revealing their raw data (for example, the Fedbully model [13]).

IV. Conclusion

This study proposed the VirtuGuard, an ethically aligned AI framework for cyberbullying detection and analysis. It integrates a cyberbullying detection AI, a knowledge graph building block, and a data analysis module into the system to provide advanced and consolidated insights based on detected cyberbullying cases and traits as well as external sources on ethics and regulations. This system functions not only as a protection/security layer in web applications for detecting and filtering cyberbullying but, more importantly, operates as an intelligent agent capable of assimilating pertinent information. It analyzes bullying patterns and their correlations with existing regulations and ethics rules to deliver distilled insights and recommendations regarding cyberbullying. The next step is to perform a comprehensive evaluation of the entire system with online data and implement privacy-preserving mechanisms to further resolve potential privacy leaks during model training and system operation.

References

[1] "Kid actions program," https://www.kidactions.eu/2022/08/04/artificial-intelligence/, 2022, [Online; accessed 14-December-2023].

[2] "Artificial intelligence act deal on comprehensive rules for trustworthy ai," https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai/, 2022, [Online; accessed 14-December-2023].

[3] B. Goldfeder and I. Griva, "Explaining cyberbullying trait detection through high accuracy transformer ensemble," in *2023 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2023, pp. 259–261.

[4] J. Wang, K. Fu, and C.-T. Lu, "Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 1699–1708.

[5] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839–1852, 2023.

[6] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury, "A human-centered systematic literature review of cyberbullying detection algorithms," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–34, 2021.

[7] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. V. Simão, and I. Trancoso, "Automatic cyberbullying detection: A systematic review," *Computers in Human Behavior*, vol. 93, pp. 333–345, 2019.

[8] Y. Zhang, M. Saberi, M. Wang, and E. Chang, "K3S: Knowledge-driven solution support system," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9873–9874.

[9] J. Zhao, T. Wen, H. Jahanshahi, and K. H. Cheong, "The random walk-based gravity model to identify influential nodes in complex networks," *Information Sciences*, vol. 609, pp. 1706–1720, 2022.

[10] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[11] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 59–68.

[12] Z. Waseem, "Are you a racist or am I seeing things? annotator influence on hate speech detection on twitter," in *Proceedings of the first workshop on NLP and computational social science*, 2016, pp. 138–142.

[13] N. P. Shetty, B. Muniyal, A. Priyanshu, and V. R. Das, "Fedbully: A cross-device federated approach for privacy enabled cyber bullying detection using sentence encoders," *Journal of Cyber Security and Mobility*, pp. 465–496, 2023.