International Journal of Information Technology (IJIT) Vol. 1, No. 3, January 2024.

MAISON: A MODEL FOR EFFECTIVE HYBRID MANAGEMENT OF CYBERSECURITY AND CYBER-TRUST

Isaak Babaev[1,2], Todd Packer[2], Mehdi Ghayoumi[1,2], Kambiz Ghazinour[1,2]

[1]Department of Cybersecurity, SUNY, Canton, NY, USA.

[2]The Advanced Information Security and Privacy Lab, Suny, Canton, NY, USA.

ABSTRACT

This paper introduces MAISON, an innovative model designed to combat cyberbullying on social media platforms. Addressing the challenge of anonymity that facilitates such behavior, MAISON integrates advanced user identification policies and employs AI-driven detection mechanisms to effectively identify and mitigate cyberbullying incidents. This approach goes beyond traditional measures, suggesting a combination of technological enhancements and policy reforms, including the use of facial motion vector detection to deter anonymous account creation for malicious purposes. The model emphasizes a holistic strategy, focusing on victim support and resilience, while advocating for robust measures against policy evasion. By aligning with emerging legal frameworks and societal demands for safer digital spaces, MAISON represents a comprehensive solution aimed at reducing both cyberbullying and its offline counterparts, thereby fostering a safer and more responsible online environment.

KEYWORDS

Cyberbullying, Privacy, Security, Society.

1. INTRODUCTION

In the evolving landscape of social media, cyberbullying has emerged as a pervasive issue, exacerbated by the veil of anonymity that digital platforms provide. Defined by Sabela et al. [7] as ’willful and repeated harm inflicted through the use of electronic devices,’ cyberbullying presents unique challenges distinct from traditional bullying. Its impact extends beyond the digital realm, often correlating with severe emotional problems, delinquency, and offline violence among individuals. The prevalence of cyberbullying, its overlap with offline bullying, and the general tendency to under-report such incidents signify a pressing need for effective intervention. strategies. This paper introduces the MAISON model, a comprehensive approach designed to identify, prevent, and mitigate cyberbullying on social media. MAISON advocates for a blend of technological innovation and policy reform, aiming to dismantle the anonymity that emboldens cyberbullies and to foster a culture of accountability and support within online communities. In this introduction, we set the stage for a detailed exploration of cyberbullying’s nuances, its sociological evolution, and the necessity for platforms to play a central role in combating it.

We underscore the urgency of adopting multifaceted solutions that resonate with the changing dynamics of online interaction and legal frameworks, such as Britain's Online Safety Bill. Through MAISON, we propose a model that not only addresses the immediate challenges of cyberbullying but also contributes to the broader discourse on creating safer digital environments.

2 CYBERBULLYING: TYPES, PERSPECTIVES, AND COPING MECHANISMS

Cyberbullying manifests in various forms, each with distinct characteristics and impacts. Researchers [8] categorize it into types such as flaming, harassment, denigration, and cyberstalking, among others. These forms share common traits of aggression, power imbalance, and the potential for recurrence, but each presents unique challenges in identification and intervention. Understanding youth perspectives on cyberbullying is crucial for effective response strategies. Studies [6] and [2] reveal gender differences in reporting and perceptions of cyberbullying, highlighting the complex dynamics of victimization and under-reporting. These insights emphasize the need for nuanced approaches to address the issue. Coping mechanisms play a pivotal role in mitigating the effects of cyberbullying. Parris et al. [5] categorize these strategies into reactive and preventive measures. Reactive coping includes avoidance, acceptance, and seeking social support, while preventive coping focuses on direct communication and enhancing online security. These strategies underscore the importance of creating supportive environments that empower victims and encourage proactive measures against cyberbullying.This section explores the multifaceted nature of cyberbullying, aiming to provide a comprehensive understanding of its forms, the affected individuals' perspectives, and effective coping mechanisms. Such understanding is essential for developing and implementing targeted interventions to combat cyberbullying in digital spaces.

3. PROPOSED METHOD

Addressing cyberbullying necessitates innovative solutions that leverage technology and community involvement. Our proposed method integrates a name-tagging system, linking social media accounts to official identification documents like driver's licenses or passports, akin to the ID.me approach. This strategy, partly mirrored by Facebook, enhances user accountability and assists in managing victimization. A significant aspect of our approach is the employment of AI to identify and flag potential cyberbullying incidents. This system is complemented by a mechanism for community members to flag cases that AI might miss, ensuring comprehensive monitoring. Verified offenders are subject to facial motion vector detection for accurate authentication, further bolstering the integrity of the platform. This dual approach of AI and community involvement lays the groundwork for evolving policy development and refining detection algorithms. To reinforce responsible online behavior, mandatory counseling, education, and peer helper programs are proposed for all involved parties. These programs are designed to educate, rehabilitate, and foster a supportive environment, thereby reducing the negative impact of cyberbullying on social networks.

3.1. Comprehensive Cyberbullying Mitigation Strategy

This section presents a holistic approach to tackle cyberbullying on social media platforms, integrating various strategies for effective management.

1) Monitor: We propose a tripartite monitoring system that includes human monitors, AI/machine learning algorithms, and a culture of collaborative community norms. This system aims to identify and flag cyberbullying incidents early, utilizing both human insight and technological efficiency.

2) Isolate: Our model equips platforms with tools to enable victims to isolate themselves from cyberbullies. This includes blocking direct interactions and removing the perpetrator’s ability to see or reference the victim, with a policy framework to escalate to expulsion if necessary.

3) Secure: To counter potential retaliation by cyberbullies, we suggest platforms implement advanced cybersecurity measures. Creating ’safe spaces’ for victims is crucial, where their communication and financial information are shielded from malicious actors.

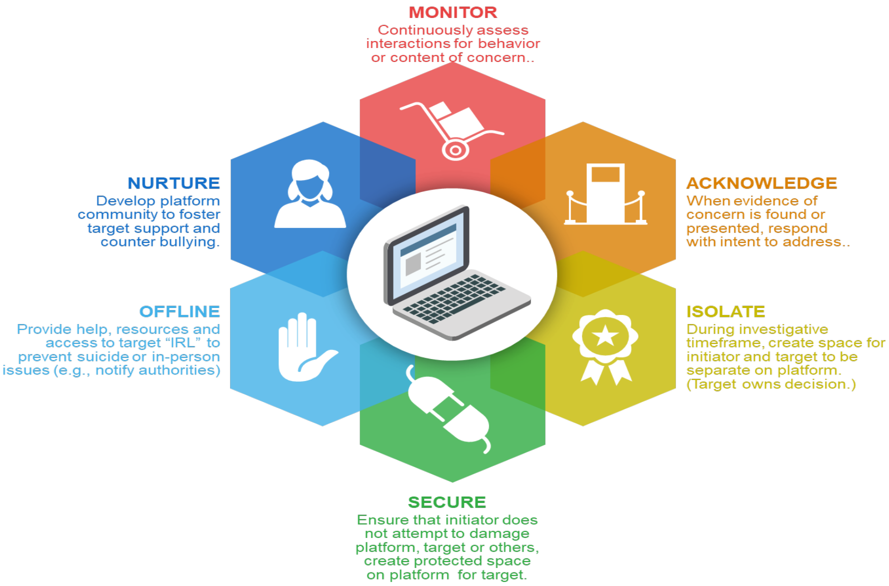
MONITOR
Continuously assess interactions for behavior or content of concern..
ACKNOWLEDGE
When evidence of concern is found or presented, respond with intent to address..
ISOLATE
During investigative timeframe, create space for initiator and target to be separate on platform. (Target owns decision.)
SECURE
Ensure that initiator does not attempt to damage platform, target or others, create protected space on platform for target.
OFFLINE
Provide help, resources and access to target “IRL” to prevent suicide or in-person issues (e.g., notify authorities)
NURTURE
Develop platform community to foster target support and counter bullying.

Fig. 1. The M.A.I.S.O.N. model provides six components of effective cyberbullying management by social media platforms.

4) Offline Measures: Acknowledging that cyberbullying has real-world impacts, our approach includes offline interventions. This involves providing resources like in-person therapy and facilitating rapid law enforcement response when necessary.

5) Nurture: The strategy also calls for platforms to foster a supportive community, countering bullying and setting positive behavioral norms. This long-term investment aims to create a safer online environment, enhancing user engagement and platform integrity.

Each element of this strategy contributes to creating a robust defense against cyberbullying, ensuring both online safety and responsible platform use.

4. AI-BASED CYBERBULLYING DETECTION

4.1. *Machine Learning Models*

In the MAISON model, machine learning is integral to detecting cyberbullying on social media. It leverages a combination of supervised, unsupervised, and semi-supervised learning methods for a comprehensive approach. Supervised learning models are utilized to process and learn from accurately labeled datasets, enabling the algorithms to identify cyberbullying patterns effectively. These models are trained and evaluated using rigorous techniques including cross-validation, hyperparameter tuning, and performance metrics like precision, recall, F1-score, and accuracy. Unsupervised learning complements this by identifying hidden patterns and anomalies in unlabeled data, which is crucial for uncovering new forms of cyberbullying. Semi-supervised learning bridges the gap between the two, enhancing the adaptability and efficiency of the models by utilizing both labeled and unlabeled data. This multifaceted approach ensures a dynamic and effective system for cyberbullying detection, maintaining a balance between sensitivity and specificity, which is crucial for ethical and trustworthy digital interaction management.

4.1.1 Transfer Learning in the MAISON Model

Transfer learning involves utilizing a pre-trained model on a large, comprehensive dataset and then fine-tuning it with a specific, smaller dataset relevant to the task at hand—in this case, cyberbullying detection. This method can significantly enhance the model's ability to generalize and understand the nuances and context of social media communications, even with limited cyberbullying-specific data.The process of transfer learning can be mathematically represented by modifying the objective function of the learning model. The model initially trained on a general dataset D_G with parameters θ_G is fine-tuned on a cyberbullying-specific dataset D_C with parameters θ_C. The objective function during fine-tuning can be represented as [11]:

$$L(\theta_C) = \sum_{(x,y)\in D_C} l(f(x;\theta_C), y) + \lambda\|\theta_C - \theta_G\|_2^2 \tag{1}$$

where $L(\theta_C)$ is the loss function for the cyberbullying dataset, l is a suitable loss function (like cross-entropy for classification), $f(x;\theta_C)$ represents the model's prediction for an input x with parameters θ_C, y is the true label, and $\lambda\|\theta_C - \theta_G\|22$ is a regularization term that keeps the fine-tuned parameters close to the pre-trained parameters, where λ is a regularization hyperparameter.

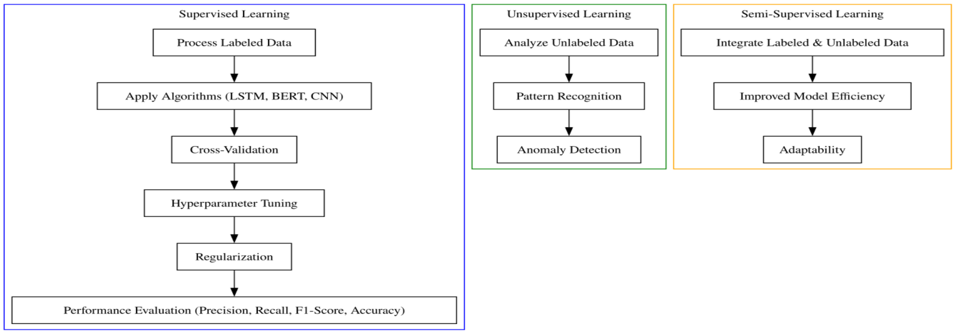

Fig. 2. Comparative Overview of Supervised, Unsupervised, and Semi-Supervised Learning Processes in Machine Learning.

5. Natural Language Processing (NLP)

In the "MAISON" model, Natural Language Processing (NLP) is employed to discern the nuanced context of social media messages, crucial for differentiating between harmful and benign communications in cyberbullying detection. NLP techniques like sentiment analysis and intent recognition analyze the tone, language, and underlying intent of textual content, allowing for the identification of negative or aggressive sentiments often associated with cyberbullying. Advanced NLP models, such as deep learning-based transformers, parse through large volumes of text to understand complex language patterns, sarcasm, and subtle cues that indicate bullying. This is complemented by intent recognition algorithms that discern the purpose behind messages, distinguishing between jokes, casual conversations, and genuine threats or harassment. By integrating these NLP capabilities, the model gains a sophisticated understanding of digital communication nuances, enabling it to effectively identify and mitigate cyberbullying while minimizing false positives and respecting the diversity of online interactions[12].

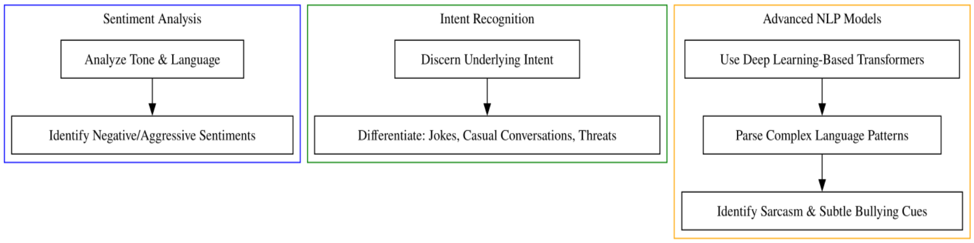
Sentiment Analysis
Analyze Tone & Language
Identify Negative/Aggressive Sentiments
Intent Recognition
Discern Underlying Intent
Differentiate: Jokes, Casual Conversations, Threats
Advanced NLP Models
Use Deep Learning-Based Transformers
Parse Complex Language Patterns
Identify Sarcasm & Subtle Bullying Cues

Fig 3. Exploring NLP Applications: Sentiment Analysis, Intent Recognition, and Advanced NLP Models.

5.1 Approaches in NLP

- **Sentiment Analysis**: Utilizing advanced sentiment analysis algorithms to identify negative sentiments in social media language, a key indicator of potential cyberbullying.
- **Intent Recognition**: Implementing machine learning techniques to discern the intent behind messages, thereby distinguishing between humor, casual talk, and potential threats.
- **Advanced NLP Models**: Leveraging deep learning-based transformers for interpreting complex language nuances, including sarcasm and subtle indicators of bullying, which are often challenging to detect with traditional models.

The mathematical formulation for NLP in cyberbullying detection is as follows. Consider a function *f* representing the NLP model that takes a textual input *x* and outputs a classification *y*. The objective function to be minimized can be represented as:

$$L(f) = \sum_{(x,y)\in D} loss(f(x), y) \qquad (2)$$

where *D* is the dataset containing pairs of textual input and their corresponding labels (indicating cyberbullying or not), and *loss* is a suitable loss function, such as cross-entropy, which quantifies the difference between the predicted classification and the actual label.

6. Discussion

The current anonymity on social media platforms significantly contributes to the prevalence of cyberbullying, an issue often underestimated and under-reported. To effectively combat this, implementing stricter user identification policies is crucial. Enhancing community self-moderation and employing AI for cyberbullying detection will not only address cyberbullying more effectively but also potentially reduce traditional bullying due to their overlap[13],[14]. Central to this strategy are counseling programs, including both individual and group sessions, designed to equip involved parties with necessary emotional coping and trauma management skills. These programs are mandatory for re-accessing social media platforms, ensuring that participants are better prepared to engage in healthier online interactions. Legal and societal pressures, as seen in initiatives like Britain's Online Safety Bill, are pushing social media networks towards a more proactive role in cyberbullying prevention. Implementing user-tagging methods will increase online accountability, and counseling will address the root causes of bullying behavior. Group sessions moderated by professionals can effectively resolve disputes and misunderstandings. Community self-moderation is particularly beneficial for those hesitant to report cyberbullying, creating a supportive online environment. Including family members in educational and therapy sessions provides additional support networks[15]. For repeat offenders, interventions may include school official involvement or account blocking, underlining the significance of cyberbullying consequences and enhancing the safety of online platforms.

7. Conclusion

This paper has explored the critical role of advanced technologies and methodologies in combating the pervasive issue of cyberbullying on social media platforms. The integration of AI, particularly through the MAISON model, has shown significant promise in detecting and addressing cyberbullying. By utilizing a blend of supervised, unsupervised, and semi-supervised learning, the model offers a comprehensive approach to identifying cyberbullying patterns. Moreover, the incorporation of natural language processing (NLP) techniques like sentiment analysis and intent recognition provides a deeper understanding of the nuances and contexts of social media interactions. The discussion emphasized the necessity of not only technological solutions but also policy and community-driven initiatives. Stricter user identification policies, enhanced community self-moderation, and the implementation of counseling programs are pivotal in creating safer online environments. The importance of legal frameworks, like Britain's Online Safety Bill, and the role of societal pressure in prompting proactive measures from social media networks were highlighted as key factors in this multifaceted approach. ;/'In conclusion, the fight against cyberbullying requires a synergistic combination of technological innovation, policy reform, and community involvement. AI and NLP offer powerful tools for detection and analysis, but their effectiveness is maximized when coupled with comprehensive counseling programs, user accountability measures, and supportive community practices.

As social media continues to evolve, it is imperative that our methods for ensuring online safety and wellbeing advance in tandem, fostering environments where positive, respectful, and healthy interactions are the norm.

REFERENCES

[1] K. Varjas, J. Talley, J. Meyers, L. Parris, and H. Cutts, “High school students’ perceptions of motivations for cyberbullying: An exploratory study,” in Western Journal of Emergency Medicine, vol. 11, no. 3, 2010, p. 269.

[2] P. Agatston, R. Kowalski, and S. Limber, “Youth views on cyberbullying,” in Cyberbullying prevention and response: Expert perspectives, 2012, pp. 57-71.

[3] S. Davis and C. Nixon, “Empowering bystanders,” in Cyberbullying prevention and response: Expert perspectives, 2012, pp. 93-109.

[4] J. Neves and L. de Oliveira Pinheiro, “Cyberbullying: A sociological approach,” in Ethical Impact of Technological Advancements and Applications in Society, IGI Global, 2012, pp. 132-142.

[5] L. Parris, K. Varjas, J. Meyers, and H. Cutts, “High school students’ perceptions of coping with cyberbullying,” in Youth and society, vol. 44, no. 2, 2012, pp. 284-306.

[6] J. L. Pettalia, E. Levin, and J. Dickinson, “Cyberbullying: Eliciting harm without consequence,” in Computers in human behavior, vol. 29, no. 6, 2013, pp. 2758-2765.

[7] R. A. Sabella, J. W. Patchin, and S. Hinduja, “Cyberbullying myths and realities,” in Computers in Human Behavior, vol. 29, no. 6, 2013, pp. 2703-2711.

[8] S. Bauman, “Types of cyberbullying,” in Cyberbullying: What Counselors Need to Know, 2015, pp. 53-58.

[9] T. Milosevic, “Social media companies’ cyberbullying policies,” in International Journal of Communication, vol. 10, 2016, p. 22.

[10] L. Wu, M. Majedi, K. Ghazinour, and K. Barker, “Analysis of social networking privacy policies,” in Proceedings of the 2010 EDBT/ICDT Workshops (EDBT ’10), Association for Computing Machinery, New York, NY, USA, 2010, Article 32, pp. 1–5. [Online]. Available: https://doi.org/10.1145/1754239.1754275

[11] M. Ghayoumi, Deep Learning in Practice, Chapman and Hall/CRC, 1st edition, 2021. ISBN-13: 978-0367458621.

[12] M. Ghayoumi, Generative Adversarial Networks in Practice, Chapman and Hall/CRC, 1st edition, 2023.

[13] K. Ghazinour and M. Ghayoumi, "An autonomous model to enforce security policies based on user's behavior," in *2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS)*, 2015, pp. 95-99.

[14] K. Ghazinour and M. Ghayoumi, "Dynamic Modeling for Representing Access Control Policies Effect," *arXiv preprint arXiv:1505.08154*, May 29, 2015.

[15] M. Ghayoumi and K. Ghazinour, "An adaptive fuzzy multimodal biometric system for identification and verification," in *2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS)*, 2015, pp. 137-141.