Predictive Modelling for Mitigating Cyber-Violence by Leveraging AI for Proactive Intervention

Lela Mirtskhulava

Department of Computer Science
Ivane Javakhishvili Tbilisi State University
Tbilisi, Georgia
lela.mirtskhulava@tsu.ge

***Abstract* - The facilitation of digital technologies increased and hardened cyber-violence which has become a significant societal concern and encompassing a wide range of harmful behaviors occurring in online environments. The given work explores the potential of predictive modeling powered by artificial intelligence (AI) to forecast future instances of cyber-violence. By analyzing historical datasets and identifying key factors contributing to online aggression, predictive models can facilitate proactive measures aimed at mitigating the risk of cyber-violence. In the given work, we discuss the methodology, challenges, and implications of employing predictive modeling in the prevention of cyber-violence. Cyber-violence forms such as cyberbullying, online harassment, hate speech, and digital abuse, pose significant threats to individuals' well-being and online communities' safety. Traditional reactive approaches to combating cyber-violence often fall short of addressing the root causes and preventing further occurrences. In response to that, predictive modeling with AI technologies comes into play for anticipating and mitigating cyber-violence before it escalates. The proposed predictive modeling involves the application of statistical algorithms and machine learning techniques to analyze historical data and predict or forecast future events. Predictive modeling aims to identify patterns, trends, and risk factors associated with online aggression in the context of cyber-violence prevention by leveraging vast datasets containing information about past instances of cyber-violence, machine learning algorithms can discern underlying factors contributing to such behaviors.**

Keywords - predictive Modelling, AI, Cyber-violence, machine learning

I. INTRODUCTION

In the digital landscape, cybersecurity has been determined as a critical concern for individuals, governments, businesses, and other sectors. Along with the complexities of cyberspace, another important and overwhelming issue called cybersecurity has also come to the forefront. Cyber-violence is treated as a range of harmful behaviors perpetrated in online space. We can differ the types of cyber-violence such as cyberbullying, online harassment, hate speech, and digital abuse. The intersection and interconnection of cybersecurity and cyber-violence are presenting unique challenges requiring innovative solutions to ensure and address the safety and security of individuals and communities in the digital world [1-5].

The given paper explores the interconnection between cybersecurity and cyber-violence and examines the various methods in which they intersect and influence one another. We will delve into factors underlying cyber-violence, its impact on individuals and societies, and the role of cybersecurity measures to mitigate its effects.

Cyber-violence poses significant threats to individuals' psychological personal safety, well-being and digital rights. Victims of cyber-violence (cyberbullying and online harassment) often experience emotional distress, social isolation, and also physical harm. The forms of cyber-violence like hate speech and digital abuse can perpetuate discrimination, and exacerbate societal tensions undermining democratic values. In terms of cybersecurity, effective measures are essential for protecting individuals' data and privacy and for safeguarding their freedoms and fundamental rights in the online space [6-8].

In response to this challenging societal concern, there is a growing interest in leveraging artificial intelligence (AI) for proactive intervention in cyber-violence. With the power of AI-driven predictive modeling, it becomes possible to mitigate instances of online aggression and avoid escalation, thereby fostering safer and more inclusive online communities.

The given paper aims to explore the potential of AI-powered proactive intervention in cyber-violence prevention and to examine the methodology, challenges, and implications of this proposed innovative approach. Analyzing historical datasets and identifying key factors contributing to cyber violence, AI-driven predictive models can facilitate early identification of individuals being at risk and communities, enabling proactive interventions and implementing the mechanisms [9-11].

By leveraging vast datasets of past instances of cyber-violence, AI algorithms can discern underlying patterns, and risk factors associated with online aggression forms. On the other hand, the adoption of AI for proactive intervention in cyber-violence prevention counts its challenges. The main challenges we meet in ethical considerations regarding data privacy, model transparency, and algorithmic bias are explored to ensure the responsible and equitable use of AI

mechanisms. Interdisciplinary collaboration and ethical innovation will lead us to harness the transformative potential of AI in creating safer and more resilient online spaces for everyone.

IoT devices such as cameras, motion sensors, and microphones can be deployed in public spaces or sensitive areas to monitor for signs of cyber-violence, such as physical altercations or verbal harassment. These devices can feed data to centralized systems for real-time monitoring and intervention.

IoT sensors integrated into smartphones or computing devices can collect data on users' online behavior, physical movements, and environmental context. Analyzing this data using machine learning algorithms can help identify patterns indicative of heightened risk for cyber-violence perpetration or victimization, enabling targeted interventions or support.

IoT infrastructure can support the development of automated crisis response systems for addressing cyber-violence incidents. Integration with emergency services, social support networks, and community organizations can enable swift and coordinated responses to reports of online harassment, hate speech, or other forms of cyber-violence.

Privacy and Security Considerations:

It's crucial to address privacy and security concerns associated with IoT deployment in cyber-violence mitigation efforts. Safeguards should be implemented to protect individuals' sensitive data collected by IoT devices, and measures should be taken to prevent unauthorized access or misuse of IoT infrastructure for surveillance or control purposes.

II. RELATED WORK

Cyber-violence is facilitated by using digital products and devices. The rising use of interconnected devices i.e. Internet of Things (IoT) devices provides benefits that greatly improve quality of life and increase the efficiency of certain tasks. On the other hand, IoT devices might collect and retain mass amounts of data and metadata on individuals and share them with a variety of parties, who may be able to extract data on where these individuals are, what they are doing or saying, and perhaps even capture imagery and videos of them. IoT security needs more attention due to security breaches increasing cybercrime threats. Our early research underscored NTRU's suitability for safeguarding IoT environments and protecting sensitive data transmitted over networks [12].

Another approach in securing IoT to maintain data integrity and protect individuals' identities. A systematic approach is essential for monitoring and mitigating potential threats to IoT security. Encryption is a fundamental requirement for ensuring secure communication within IoT networks. Key exchange is pivotal in securing information exchange within IoT networks. Neural networks offer an effective strategy through synchronization using the Hebbian learning rule to balance weights, providing a cryptographic key-exchange protocol. A significant advantage of this process is the extended time required for an attacker to guess the generated key [13].

AI gives us many opportunities to detect cyber-violence and its forms. Machine learning (ML) and deep learning (DL) algorithms can detect early signs of cyber-violence. AI prevents cyber-violence from spreading further. AI makes it possible to offer victims personalized post-cyber-violence treatment. Below we describe various techniques and approaches utilized in identifying abusive language and hostile interactions in online platforms.

A. Predictive Analytics for Online Harassment Detection

Early research explored the application of predictive analytics to detect and mitigate online harassment. Machine learning and natural language processing (NLP) have been applied to analyze textual data to identify abusive language and hostile interactions in online platforms [14].

B. Sentiment Analysis and Cyberbullying Detection

Early studies have explored sentiment analysis algorithms to identify indicators of cyberbullying and negative interactions in social media and online forums. Researchers have developed models to classify and flag potentially harmful content by analyzing linguistic cues and emotional patterns [15].

C. Machine Learning for Content Moderation

Machine learning algorithms have gained attention for content moderation in online platforms. These algorithms analyze user-generated content to detect and filter out harmful or abusive material, thereby reducing cyber-violence dissemination [16].

D. Behavioral Analytics for Risk Assessment

Behavioral analytics have been explored to assess the risk of individuals facing cyber-violence. Going through analysis of patterns of online behavior, such as the frequency of aggressive interactions or engagement with contentious topics, predictive models can identify individuals at heightened risk of perpetrating or experiencing cyber violence [17-20].

E. Ethical Considerations in AI-driven Intervention

Data Scientists have highlighted the ethical implications of using AI-driven interventions to address cyber-violence. Concerns include potential biases in predictive models, privacy implications, and the balance between intervention and free speech. Ethical frameworks and guidelines have been proposed to ensure responsible and transparent deployment of AI technologies in this context [21-23].

III. THE PROPOSED MODEL

The goal is to develop a system capable of automatically detecting cyber-violence forms in text messages. Cyber-violence involves the use of digital communication platforms to harass, intimidate, or harm others. Detecting cyber-violence forms involves analyzing text data for signs of harmful or abusive language. One common approach is to use machine learning techniques to classify text as either indicative of cyber-violence or not.

We developed Python code using a basic machine learning classifier, specifically a Support Vector Machine (SVM) with TF-IDF vectorization for feature extraction. It can preprocess the text data by converting it to lowercase, punctuation, stopwords and removing numbers. It evaluates the model's accuracy and prints a classification report.

The decision function for SVM is represented as:

$$f(x)=\text{sign}(w \cdot x+b) \quad (1)$$

where:

f(x) is the decision function that predicts the class label of input x
w is the weight vector
x is the input feature vector
b is the bias term
sign(·) is the sign function that returns -1 if the argument is negative, 0 if it's zero, and 1 if it's positive.

The proposed code assumes the use of a labeled dataset for training a robust model where each message is labeled as either cyber-violence (1) or not (0).

TF-IDF (term frequency-inverse document frequency) is a measure, used in the domains of information retrieval (IR) and machine learning, that can quantify the importance of string representations such as words, phrases, lemmas, etc.

IV. DETAILED DESCRIPTION OF THE GIVEN METHODOLOGY

A. Data Collection

Data Collection: The first step in developing a machine learning model is to collect a dataset. We're using a dataset containing text messages labeled as either one of the cyber-violence forms like cyberbullying or not cyberbullying. The dataset consists of a text message and its corresponding label (0 for not cyberbullying, 1 for cyberbullying). Building a high-quality dataset is crucial for the success of the model.

B. Data Processing

Text Cleaning: The text data often contains noise such as punctuation, digits, and special characters. Text cleaning means removing these unwanted elements from the text.

Tokenization: Tokenization is the procedure of breaking down text into smaller units, such as n-grams (sequences of n words). In the given mode, we tokenize the text into individual words.

Stopword Removal: Removing stopwords can help reduce the dimensionality of the data and focus on more meaningful words. Stopwords are common words that often do not carry significant meaning (e.g., "the", "is", "and").

C. Normalization

Normalization involves converting text to a consistent format, such as converting all letters to lowercase.

Feature Extraction: Machine learning algorithms require numerical input data. In this model, we use the Term Frequency-Inverse Document Frequency (TF-IDF) technique for feature extraction. TF-IDF measures the importance of each word in a document relative to a collection of documents. It assigns higher weights to words that are frequent in the document but rare in the entire corpus.

Model Selection and Training: SVM is a popular supervised learning algorithm used for classification tasks. It works by finding the hyperplane that best separates the classes in the feature space.

Training: We split the dataset into training and testing sets. The training set is used to train the SVM model on the features extracted from the text data.

D. Model Evaluation

Testing: We evaluated the trained model on the testing set to assess its performance.

Accuracy: Accuracy measures the proportion of correctly classified instances out of all instances in the testing set.

Classification: The classification process provides precision, recall, F1-score, and support for each class. Precision measures the proportion of true positive predictions out of all positive predictions.

Deployment: Once the model is trained and evaluated satisfactorily, it can be deployed in a real-world application where it can automatically detect cyberbullying in text messages (Fig.1).

```
messages = [
    ("I hate you, you're so stupid!", 1),   # Cyberbullying message (1)
    ("You're awesome, keep it up!", 0),     # Non-cyberbullying message (0)
    ("I'm going to hurt you!", 1),          # Cyberbullying message (1)
    ("I disagree with you, but that's okay.", 0),  # Non-cyberbullying message (0)
```

Fig.1. Message in Python code

V. CONCLUSION

In the given paper, we have proposed a comprehensive methodology for the development of a system detecting cyber-violence forms leveraging Python code and machine learning methodologies. We have demonstrated the efficacy of Python as a versatile tool in structuring robust cyberbullying detection model by systematically addressing each stage, from data collection to potential deployment.

Through data collection and preprocessing, we used a dataset essential for training and testing our proposed detection system. Employing Python libraries such as scikit-learn and NLTK facilitated seamless data manipulation, enabling effective feature extraction and model training. We have used the TF-IDF technique to enhance our model's capability to discern significant patterns within text data, a critical aspect of cyber-violence detection. Evaluation metrics, including accuracy and classification reports, provided insightful assessments of the model's performance. These metrics underscored the system's capability to discriminate between cyberbullying and non-cyberbullying

content, laying the groundwork for future enhancements and real-world deployment.

This research contributes to the growing body of literature aimed at mitigating the adverse impacts of cyberbullying in digital communication platforms. By harnessing Python's versatility and machine learning techniques, we present a viable approach to creating proactive solutions for fostering safe and secure online environments and promoting positive digital interactions. As technology evolves, continued exploration and refinement of such methodologies hold promise in advancing cyberbullying prevention efforts and enhancing the well-being of digital communities.

REFERENCES

[1] GREVIO (Council of Europe Group of Experts on Action against Violence against Women and Domestic Violence), General Recommendation No1 on the digital dimension of violence against women, Council of Europe, Strasbourg, 20 October 2021 (https://rm.coe.int/grevio-rec-no-on-digital-violence-against-women/1680a49147).

[2] Van der Wilk, A. Cyber Violence and Hate Speech Online against Women, European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs, Brussels, 2018.(https://www.europarl.europa.eu/RegData/etudes/STUD/2018/604979/IPOL_STU(2018)604979_EN.pdf).

[3] Lomba, N., Navarra, C., and Fernandes, M., Combating Gender-based Violence: Cyber violence – European added value assessment, European Parliamentary Research Service, Brussels, 2021.(https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662621/EPRS_STU(2021)662621_EN.pdf).

[4] EIGE (European Institute for Gender Equality), Gender Equality Index Report, Vilnius https://eige.europa.eu/publications/gender-equality-index-2020-report), 2020.

[5] (https://rm.coe.int/grevio-rec-no-on-digital-violence-against-women/1680a49147).

[6] EIGE , Gender equality and digitalization in the European Union, Vilnius (https://eige.europa.eu/publications/gender-equality-and-digitalisation-european-union), 2018.

[7] Council of Europe: YOUR DIGITAL RIGHTS IN BRIEF. https://rm.coe.int/1680301b6e

[8] Internet Freedom and Digital Rights in Georgia: Systemic Challenges.

[9] https://idfi.ge/en/internet_freedom_and_digital_rights_in_georgia Internet Freedom in Armenia and Execution of Basic Human Rights in Online Freedom.

[10] https://mediainitiatives.am/wp-content/uploads/2018/03/Internet-Freedom-Research-Report-2017-in-English.pdf

[11] The internet as a human right. https://www.brookings.edu/articles/the-internet-as-a-human-right/

[12] L. Mirtskhulava, L. Globa, N. Meshveliani and N. Gulua, "Cryptanalysis of Internet of Things (IoT) Wireless Technology," 2019 International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), Odessa, Ukraine, 2019, pp. 1-4, doi: 10.1109/UkrMiCo47782.2019.9165363.

[13] L Mirtskhulava, N Gulua, N Meshveliani. IoT SECURITY ANALYSIS USING NEURAL KEY EXCHANGE PROTOCOL Georgian Electronic Scientific Journal, ISSN 1512-1232 2 (57).

[14] Dadvar, M., & de Jong, F. (2011). Cyberbullying detection: a step toward a safer internet. In Proceedings of the Workshop on Current Trends in News Information Retrieval co-located with 35th European Conference on Information Retrieval (ECIR 2013.

[15] Chatzakou, D., Kourtellis, N., Blackburn, J., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1, pp. 27-36).

[16] Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2017). Prominent features of rumor propagation in online social media. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (pp. 595-602).

[17] Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2), 223-242.

[18] Ribeiro, F. N., Santos, L. F., Macdonald, C., & Cerqueira, R. (2018). Characterizing and detecting hateful users on Twitter. ACM Transactions on Internet Technology (TOIT), 18(3), 1-20.

[19] Sood, S., & Anto, L. (2017). Cyberbullying Detection on Twitter Using Sentiment Analysis. In Proceedings of the 18th International Conference on Electronic Commerce: e-Business in Smart Applications (pp. 1-6).

[20] Wimalasuriya, D. C., & Bendersky, M. (2010). Cyberbullying detection with machine learning algorithms. In 2010 IEEE Second International Conference on Social Computing (pp. 409-416). IEEE.

[21] Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... & Vakali, A. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In Twelfth International AAAI Conference on Web and Social Media.

[22] Zhang, A. X., & Dixon, L. (2019). The Effects of Data Preprocessing on Hate Speech Detection on Twitter. In Proceedings of the Third Workshop on Abusive Language Online (pp. 134-140).

[23] Van Hee, C., Bontcheva, K., & Stringhini, G. (2020). Automated hate speech detection and the problem of offensive language. IEEE Intelligent Systems, 35(4), 80-89.