CYBERBULLYING DETECTION ON SOCIAL NETWORKING USING NLP ALGORITHM

Zibeon Samuel Moses
Department of Computer Scienceand Engineering
Sathyabama Institute of Scienceand Technology
Chennai, India
Zibeonmoses@gmail.com

Sandra Maria George
Department of Computer Scienceand Engineering
Sathyabama Institute of Scienceand Technology
Chennai, India
Sandramariageorge03@gmail.com

Mrs. Scinthia Clarinda,B.E.,M.E.,
Assistant Professor, Department of Computer Science and Engineering
Sathyabama Institute of Scienceand Technology
Chennai, India

Abstract - Automatic identification of cyberbullying is a problem that is gaining traction, especially in the Machine Learning areas. Not only is it complicated, but it has also become a pressing necessity, considering how social media has become an integral part of adolescents' lives and how serious the impacts of cyberbullying and online harassment can be, particularly among teenagers. This paper contains a systematic literature review of modern strategies, machine learning methods, and technical means for detecting cyberbullying and the aggressive command of an individual in the information space of the Internet. We undertake an in-depth review of 13 papers from four scientific databases. The article provides an overview of scientific literature to analyze the problem of cyberbullying detection from the point of view of machine learning and natural language processing. In this review, we consider a cyberbullying detection framework on social media platforms, which includes data collection, data processing, feature selection, feature extraction, and the application of machine learning to classify whether texts contain cyberbullying or not. This article seeks to guide future research on this topic toward a more consistent perspective with the phenomenon's description and depiction, allowing future solutions to be more practical and effective.

I. INTRODUCTION

The modern space of everyday communication is characterized by a new striking feature-its spread into the virtual world. If for modern adults communication skills using e-mail, instant messages and chats are an addition to the already acquired skills of live communication, then modern children and adolescents master both of these skills almost simultaneously. As for adolescents, we can say This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. that the process of socialization is largely moving to the Internet–along with acquaintances, reference groups, the development of various social roles and norms. All those communicative processes that occur in ordinary sociophysical space are "duplicated", sometimes intensified, and sometimes compensated by virtual communication, but in any case acquire new features. Although historically virtual existence is clearly secondary in relation to the real one, one can expect a reverse impact and transfer of communicative situations and rules common on the Internet into the "real" space of communication. With the development of information technologies, significant changes have taken place in the life of a modern teenager: virtual reality has appeared, in which communication and interpersonal relationships are moving to a new, unfamiliar level for them. Bullying becomes more dangerous for an individual since it can be carried out using Internet technologies. For the first time, the definition of "cyberbullying" was given by Bill Belsey. In his opinion, cyberbullying is the use of information and communication technologies, for example, e-mail, mobile phone, personal Internet sites, for intentional, repeated, and hostile behavior of a person or group aimed at insulting other people. Bullying98+ on the Internet can be carried out 24 h a day, 7 days a week, leaving no chance to feel protected, messages and comments can come unexpectedly, at any time–this has a strong psychological impact on a teenager. There is also anonymity on the Internet, thanks to which a teenager may not even suspect what kind of person is bullying him, which can cause him even more fear. Unlike physical violence, the consequences of emotional violence in the long term affect psychological health. Therefore, our goal is to study the phenomenon of cyberbullying as a form of suppression of adolescent personality and determine the content of cyberbullying prevention using machine learning technology, which involves setting the following tasks: types of cyberbullying in online content; identify methods and tools for automatic detection of cyberbullying and hateful expressions; consider open datasets for training machine learning models for automatic detection of cyberbullying; highlight the state-of-the-art methods and analyze the futuretrends. This paper is organized as follows: Section 2 explains the literature review method. Section 3 reviews cyberbullying, digital drama, hate speech problems and describes types of cyberbullying.

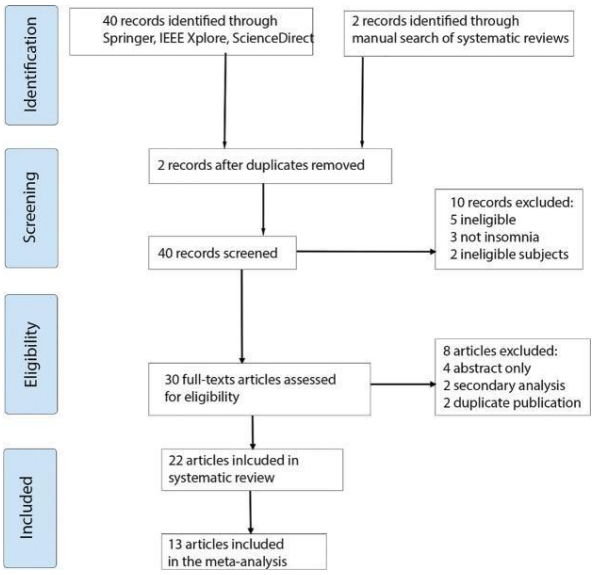
Section 4 reviews research papers in cyberbullying detection area. In this section, we describe each stage of cyberbullying detection on social media from data collection to text classification. Section 5 discusses existing problems and research challenges. Finally, in the last section, we conclude our review.

II. LITERATURE SURVEY

As a constructive mode of information sharing, collaboration and communication, social media platforms offer users with limitless opportunities. The same hypermedia can be transposed into a synthetic and toxic milieu that provides ananonymous, destructive pedestal for online bullying and harassment. Automatic cyberbullying detection on social media using synthetic or real-world datasets is one of a proverbial natural language processing problem. Analyzing a given text requires capturing the existent semantics, syntactic and spatial relationships. Learning representative features automatically using deep learning models efficiently captures the contextual semantics and word order arrangement to build robust and superlative predictive models. This work puts forward a hybrid model, Bi-GRU-Attention-CapsNet (Bi-GAC), that benefits by learning sequential semantic representations and spatial location information using a Bi-GRU with self-attention followed by CapsNet for cyberbullying detection in the textual content of social media. The proposed Bi-GAC model is evaluated for performance using F1-score and ROC-AUC curve as metrics. The results show a superior performance to the existing techniques on the benchmark Formspring.me and My Space datasets. In comparison to the conventional models, an improvement of nearly 9% and 3% in F-score is observed for My Space and Formspring.me dataset respectively. Nearly 60% of teenagers in the USA have experienced abusive online behaviour. Identifying effective programmes to address these behaviours and promote digital citizenship is a research priority to reduce the rate of occurrence and consequential harmful effects of abusive online behaviour. To evaluate the effectiveness of a Digital Citizenship Curriculum in increasing knowledge of digital citizenship and reducing cyberbullying and online aggression among middle-schoolers in an underserved community using a free curriculum. Middle-schoolers participated in pilot implementation of a Digital Citizenship Curriculum (DCC) to evaluate its effectiveness in increasing knowledge of digital citizenship and reducing cyberbullying and online aggression. Follow up interviews were conducted to explore participants' perceptions of the curriculum. Participants demonstrated a statistically significant increase in their knowledge of digital citizenship with an increase of 2.96 in the mean score (p < .001). Paired t-tests by gender demonstrated a significant difference in pre-post assessment mean scores for girls (p < .001). Post-intervention perceptions indicate the curriculum was positively received and informative. Identifying cost-effective and resource-friendly programmes that support social-emotional learning and promote digital citizenship is crucial for underserved populations. Regions such as Appalachian Ohio often lack the resources to fund costly curriculum aimed at online aggression prevention. This study supports the implementation of the DCC and indicates the need for future research on the long-term effects of the curriculum on middle school participants.

III. PROPOSED METHODOLOGY

In this literature review, both qualitative and quantitative analysis methods were integrated and applied shows the steps of the review and the number of included and excluded articles. The collection of articles started by defining a search string. This string is composed of three search terms: "Machine Learning" and "Cyberbullying" with the logic operator "AND" in between them. Four reference databases were used: Science Direct Platform, IEEE Xplore digital library, Springer, Wiley online library. We included in our research the articles published between 2015 to 2021. In the Screening stage, we excluded ten records. In the Eligibility stage, we excluded eight records without full texts, secondary analysis, and duplicated publications. From the remaining 22 papers, we left 13 for meta-analysis.



Classification of Cyberbullying

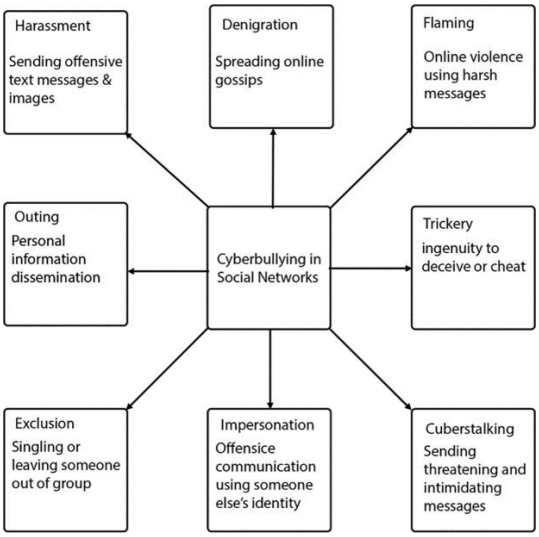
Digital drama is a new age phrase that refers to forms of abuse and violence among teens inthe technology world [9]. The National Crime Prevention Council's definition of cyber-bullying is "when the internet, cell phones or other devices are used to send or post text or images intendedto hurt or embarrass another person" [10]. StopCyberbullying.org, an expert organization dedicated to internet safety, security, and privacy, defines cyberbullying as: "a situation when a child, tween orteen is repeatedly 'tormented, threatened, harassed, humiliated,

embarrassed or otherwise targeted' byanother child or teenager using text messaging, email, instant

messaging or any other type of digital technology".

Cyberbullying can be as simple as continuing to send e-mail to someone who has said they want no further contact with the sender, but it may also include threats, sexual remarks, pejorative labels (i.e., hate speech), ganging up on victims by making them the subject of ridicule in forums, and posting false statements as fact aimed at humiliation.

Like traditional bullying, cyberbullying can be direct and indirect . Direct cyberbullying is direct attacks on a child through letters or messages. In case of indirect harassment, other people (both children and adults) are involved in the process of victim harassment, not always with their consent; the stalker can hack the victim's account and, mimicking the host, send messages from this account to the victim's friends, destroying the victim's communicative field and creating doubt about his moral qualities.



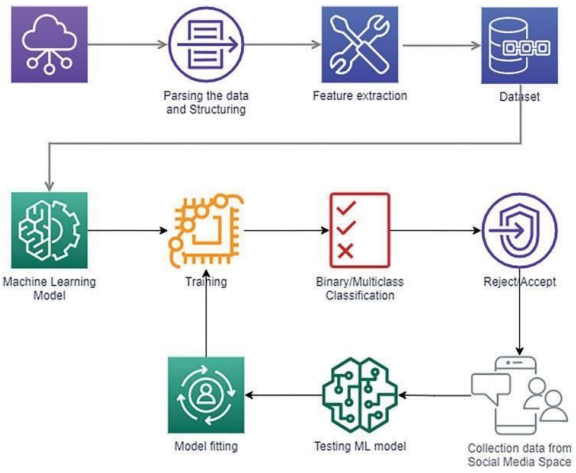
The most emotionally violent form of cyberbullying is Flaming, which begins with insults and develops into a quick emotional exchange of remarks, usually in public, less often in private correspondence. It occurs between two interlocutors with initially equal positions, but sudden aggression introduces an imbalance, which is amplified by the fact that the participant does not knowwho his opponent can attract to his side in this battle. Forum visitors, witnesses, can join one ofthe parties and develop rough correspondence, not fully understanding the original meaning of the collision and often considering the situation as a game, unlike the initiators of an aggressive dialogue. You can compare this with a "wall-to-wall" fight, where the participants do not fully understand either what was the reason for the conflict, or what is the criterion for joining comrades-in-arms to all.

Machine Learning in Cyberbullying Detection

Threats arising in the network environment naturally stimulate the development of a research apparatus for studying their factors, mechanisms, and consequences. The emotional nature of those processes that mainly determine the destructive nature of the impact of negative network phenomena on a person sets as a priority the creation of automatic analysis tools that allow assessing the severityof signs of affective states of communication participants in network communication, that is, methodsof sentiment analysis in the broad sense of the word. illustrates cyberbullying detection on social networks using machine-learning techniques. The figure describes cyberbullying detection process from data collection to cyberbullying text classification.

Data Collection

Most ML-based text classification models rely on data as a key component. Data, on the other hand, is meaningless unless it is used to derive information or implications. Training and testing datasets are chosen using data gathered from social networks. Based on observed cases (marked data), supervised models strive to give computer approaches for improving classification accuracyin specified tasks. A good model for a given task should not be confined to samples in a training set alone, and therefore should contain unlabeled actual data. The amount of data is unimportant; what matters is if the retrieved data accurately represents social media website activity. Data extracted from social media using keywords, keyphrases, or hashtags, or data extractedfrom social networks using user profiles are the two main data collection techniques in cyberbullying detection studies. The Data Gathering section highlights the problems withdifferent data collection methodologies and their impact on ML



Discussion

The goal of this research was to see whether cyberbullying could be automatically identified usingthe parameters that make up its identity and characteristics. As a result, we utilized a systematic literature review method to provide an in-depth overview of studies on automated cyberbullying

detection. In accordance with the results of the given review, we offer recommendations for future study and suggest enhancements to existing machine learning models and classifiers in automated cyberbullying detection in this section.To summarize, and per our second goal, we propose that future research take into consideration and completely disclose a set of essential information in order to enhance the quality of future datasets, models, and the performance of classifiers. It is critical to give annotators clear instructions based on the characteristics that define cyberbullying (i.e., intentionality, repetition, aggression, and peer conduct), as well as to guarantee that the annotators are specialists in the area of cyberbullying. Furthermore, data extraction from users should be acquired through peers, and users' privacy should be prioritized during this process. In addition, methods should be created to try to capture the context and nature of the players' relationships in a cyberbullying incident, since this is a critical component in identifying deliberate damage and repeated aggressions among peers.

IV. CONCLUSION

Social media is a relatively new human communication medium that has grown in popularityin recent years. Machine learning is utilized in a variety of applications, including social network analysis. This review provides a thorough overview of different applications that use machine learningtechniques to analyze social media to detect cyberbullying and online harassment. Our paper explores each step to cyberbullying detection on social media, such as data collection, data preprocessing, data preparation, feature selection and extraction, feature engineering, applying machine learning techniques, and text classification. Various academics have proposed various methods to address the problems of generic metadata architecture, threshold settings, and fragmentation in cyberbullying detection on social networks data streams. To address problems with cyberbullying categorization in social network data, the review also proposed a general metadata architecture for cyberbullying classification on social media. When compared to comparable techniques, the proposed architecture performed better across all evaluation criteria for cyberbullying and online harassment detection. In further, a more durable automated cyberbullying detection system can be developed byconsidering the problems as class imbalance data, binary and multi-classification, scalability, class imbalance data, multilingualism, threshold settings, and fragmentation.

V. RFERENCE

[1] A. Orben, "Teenagers, screens and social media: A narrative review of reviews and key studies," *Social Psychiatry and Psychiatric Epidemiology*, vol. 55, no. 4, pp. 407–414, 2020.

[2] D. Al-Sabti, A. Singh and S. Jha, "Impact of social media on society in a large and specific to teenagers," in *6th Int. Conf. on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, pp. 663–667, 2017.

[3] A. Kumar and N. Sachdeva, "A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media," *World Wide Web*, vol. 25, no. 4, pp. 1537–1550, 2022.

[4] B. Belsey, "Cyberbullying: An Emerging Threat to the «Always on» Generation," 2019. [Online].

[5] M. Boniel-Nissim and H. Sasson, "Bullying victimization and poor relationships with parents as risk factors of problematic internet use in adolescence," *Computers in Human Behavior*, vol. 88, pp. 176–183, 2018.

[6] P. K. Bender, C. Plante and D. A. Gentile, "The effects of violent media content on aggression," *Current Opinion in Psychology*, vol. 19, no. 1, pp. 104–108, 2018.

[7] Z. Munn, M. D. Peters, C. Stern, C. Tufanaru, A. McArthur *et al.*, "Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach," *BMC Medical Research Methodology*, vol. 18, no. 1, pp. 1–7, 2018.

[8] M. J. Grant and A. Booth, "A typology of reviews: An analysis of 14 review types and associated methodologies," *Health Information and Libraries Journal*, vol. 26, no. 2, pp. 91–108, 2009.

[9] M. Brandau, T. Dilley, C. Schaumleffel and L. Himawan, "Digital citizenship among appalachian middle schoolers: The common sense digital citizenship curriculum," *Health Education Journal*, vol. 81, no. 2, pp. 157–169, 2022.

[10] S. Day, K. Bussey, N. Trompeter and D. Mitchison, "The impact of teasing and bullying victimization on disordered eating and body image disturbance among adolescents: A systematic review," *Trauma, Violence and Abuse*, vol. 23, no. 3, pp. 985–1006, 2022.

[11] J. S. Hong, D. H. Kim, R. Thornberg, J. H. Kang and J. T. Morgan, "Correlates of direct and indirect forms of cyberbullying victimization involving south Korean adolescents: An ecological perspective," *Computersin Human Behavior*, vol. 87, pp. 327–336, 2018.