International Journal of Advanced Computer Technology
ISSN: 2319-7900 www.ijact.org Volume 12, Issue 4, July-August 2023

Machine Learning Methods for Detection of Bystanders: A Survey

Khushboo Thakkar, Anamika Gupta, Vibhor Mathur, Aman Tiwari
SSCBS, University of Delhi, Delhi, India.

Abstract:- *The number of users on social media networks is increasing daily due to the rising popularity of these platforms. These users share their photos, videos, daily experiences, views, and status updates on various social networking sites. While social networking sites offer great possibilities for young people to interact with others, they also expose them to unpleasant phenomena such as online harassment and abusive language, resulting in cyberbullying. Cyberbullying is a pervasive social problem that has detrimental consequences for the health and safety of its victims, including psychological distress, anti-social behaviour, and even suicide. The Bystander role plays a crucial part in minimising the impact of cyberbullying. This paper presents a review of cyberbullying content on the internet, the classification of cyberbullying categories, the categorisation of author roles (harasser, victim, bystander-defender, bystander-assistant), data sources, and machine learning techniques for detecting cyberbullying.*

Introduction

Social media platforms have evolved into remarkable tools for connecting individuals worldwide. However, as these online platforms gain popularity in the digital realm, some utilise them positively, while others engage in reprehensible actions. Cyberbullying is one concerning issue that has emerged due to the proliferation of social media platforms [1]. Cyberbullying entails using digital technology, such as smartphones, computers, and tablets, to engage in bullying behaviours. It can transpire through applications, online social media, forums, and gaming platforms where users interact, exchange content, or participate in discussions. Cyberbullying encompasses acts such as sending, uploading, or disseminating hurtful, false, derogatory content about others, including disclosing personal or private information leading to embarrassment or humiliation. Certain forms of cyberbullying are even illegal or criminal [2].

The impact of cyberbullying on youth is considerable. In the context of cyberbullying, bystanders are individuals who witness bullying incidents online, which may involve even strangers. Witnessing cyberbullying is distressing and affects bystanders as well. Bystanders hold the potential to make a positive impact in such situations by assuming various responsibilities. The presence of supportive peers can alleviate the distress and unhappiness experienced by bullied individuals. In fact, during instances of bullying, bystanders are present 80% of the time, and when they intervene, the bullying ceases in 57% of cases within 10 seconds [3].

Toxic behaviour often unfolds in the presence of bystanders. In such scenarios, bystanders can assume different roles to alter the dynamics of social situations. Bystanders play a crucial role in handling situations involving toxic behaviour. They can react in three ways: mirroring the perpetrator's toxic behaviour (inadvisable), hindering the toxic conversation and standing up for the victim (recommended), or simply observing the unfolding events. The dynamics of bystander engagement in prosocial behaviour within cyberspace in response to hate speech, cyberbullying, or trolling are intricate. This complexity arises because the presence of other internet users might lessen one's sense of responsibility to intervene, assuming that someone else will take action. However, in smaller groups, bystanders feel a stronger obligation to intervene in instances of cyberbullying [4]. Bystanders play an essential role in preventing and intervening in bullying. Their roles encompass various aspects, such as Outsiders are Individuals who observe the situation without getting involved. Defenders are

Individuals who intervene and support the victim. Reinforcers are Individuals who support and encourage bullying behaviour. Assistants are Individuals who join in bullying activities. It's important to differentiate between being a bully and supporting a bully. Thus, a bystander-reinforcer behaves as a cheerleader for the bully, while a bystander-assistant assists the bully by adopting bullying behaviour themselves [3].

Types of Cyberbullying

Cyberbullying can take various forms, necessitating a comprehensive understanding of these types for effective prevention and intervention. The following are common types of cyberbullying [9,10,11,12,13,14,15]:

1. Flaming describes rapidly escalating online discussions or conflicts involving abusive language, insults, or personal attacks. Flaming is prevalent in public online spaces like comment sections and discussion forums.
2. Harassment: Harassment involves repeatedly sending disrespectful, threatening, or offensive messages or comments to an individual. This can occur across various digital platforms, such as social media, messaging apps, and online forums.
3. Cyberstalking: Cyberstalking refers to the persistent surveillance, tracking, or harassment of an individual online. It includes unwanted and intrusive communications, continuous monitoring of online actions, and threats that induce fear or discomfort.
4. Masquerade occurs when a bully adopts a fake identity to target someone anonymously. Beyond creating a fabricated identity, the bully might impersonate another person to send malicious messages to the victim.
5. Trolling: Trolling involves intentionally provoking negative reactions from online users by posting controversial or provocative messages or comments. Trolls often aim to evoke emotional responses or disrupt discussions for their amusement or attention.
6. Denigration: Denigration involves sharing false, hurtful, or negative information about an individual to tarnish their reputation or cause emotional distress. Cyberbullies use posts, comments, or messages to belittle, ridicule, or disclose personal and humiliating details. This type of cyberbullying can spread quickly and have lasting consequences for the victim's self-esteem, relationships, and well-being.
7. Outing: Outing occurs when a bully publicly exposes personal and private information, photos, videos, or any sensitive data about someone. The victim becomes "outed" as their information becomes widely available online.
8. Exclusion: Exclusion entails intentionally singling out and excluding someone from an online group, such as a chat room or website. The group then engages in derogatory comments and harassment against the targeted individual.
9. Catfishing: Catfishing involves creating a fictitious online persona to build relationships, gain trust, and manipulate others for various purposes, including harassment, emotional manipulation, or fraud. Catfishers may use bogus profile images, fabricated biographical information, and elaborate stories to deceive victims, typically on social media or online dating platforms.
10. Dissing: Dissing refers to cyberbullying, where individuals insult or mock others online. It encompasses using offensive language, insults, or disrespectful statements that ridicule or belittle the target.
11. Trickery: Trickery involves deceiving or manipulating someone through cyberbullying to cause emotional or psychological harm. This may include manipulative hoaxes, fake friendships, dangerous online challenges, identity theft, false promises, or blackmail to exploit the victim emotionally, psychologically, or socially.
12. Fraping: The term "fraping," derived from "Facebook" and "hijacking," denotes gaining unauthorised access to someone's social media

account and posting or sharing content without consent. It entails posing as the individual and publishing messages, status updates, or other content on their behalf. Done maliciously, fraping aims to humiliate, embarrass, or harm the compromised account holder, leading to emotional distress, reputation damage, or other negative repercussions.

Impact of Cyberbullying

Children and teenagers are embracing the internet more extensively, at younger ages, and through diverse channels than ever before. This trend has given rise to a significant concern: cyberbullying [5]. Due to the prevalence of a social lifestyle that minimises face-to-face interactions, it's imperative to explore and research the domain of cyberbullying collectively. Moreover, the lack of well-established legal frameworks for cyberbullying in most countries contributes to a clouded understanding of addressing the issue. Social media has revolutionised our lifestyle and business practices by enabling real-time interactions. Despite the numerous advantages of social media, negative consequences also emerge. Among the most severe misuse of digital media is cyberbullying, wherein these platforms are wielded to incite anger, intimidation, or humiliation towards other online users. Certain individuals exploit these platforms to share distorted information, manipulate photographs, pen nasty remarks, and publish videos intended to harm or disgrace others. Cyberbullying has been demonstrated to have lasting repercussions on victims, resulting in stress, persistent anguish, sleep disorders, and even issues like hunger [6].

The psychological impact of cyberbullying on victims is substantial. They experience depression, loneliness, anxiety, and even contemplate suicide. Physical health is also severely affected, leading to headaches, insomnia, abdominal pain, eating disorders, and nausea. Studies indicate that approximately 8% of cyberbullying victims consider suicide. The self-esteem of victims is significantly diminished by such incidents, leading to feelings of isolation and suicidal ideation [7][8].

A particular study delved into the connection between cyberbullying and mental health issues among high school students. The study used logistic regression to examine the relationship between cyberbullying and students' mental well-being. The findings revealed that being a victim of cyberbullying predicts adverse mental health outcomes irrespective of ethnicity, gender, or grade level. The study suggests that being victimised by cyberbullying doubles the likelihood of engaging in drug abuse and experiencing suffering while also tripling the chances of contemplating suicide [16].

Related Work in Bystander Detection

According to current definitions, cyberbullying is described as the sharing of online content by an individual that is intolerant or harmful to a victim. To identify bullying, an annotation technique [17] was developed using criteria to identify textual features of cyberbullying, including posts by bullies and responses from victims and bystanders. Another study on cyberbullying focuses on identifying the roles of its participants [18]. They were among the first to categorise roles in a bullying scenario. Based on surveys of teenagers involved in real-life bullying situations, they defined six participant roles: victims (target of repeated harassment), bullies (initiative-taking perpetrators), assistants of the bully (encourage the bullying), reinforcers of the bully (reinforce the bullying), defenders (comfort the victim, take their side, or try to stop the bullying), and outsiders. In total, the researchers distinguish four bystander roles (assistants, reinforcers, defenders, and outsiders) in addition to the bully and victim.

Cyberbullying is not limited by language barriers. The primary objective of the research in [19] is to gain insight into the linguistic characteristics of cyberbullying. This is achieved by collecting and annotating a dataset in two phases. In the first phase, annotators assign a harmfulness score to determine whether a post is part of a cyberbullying event. If it is, annotators then classify the authors' roles into four categories (Harasser, Victim, Bystander defender, and bystander assistant). In the second phase, annotators identify fine-grained text categories related to cyberbullying, even if the post wasn't initially considered harmful. For experimental purposes, a binary classifier is developed to distinguish bully posts. By the end, a binary classifier is built for each fine-grained bullying category. Features such as bag-of-word and polarity features based on existing sentiment lexicons are implemented [20] to automatically identify cyberbullying as a binary classification problem and roles as a multiclass classification problem. This relies on supervised learning mechanisms using pre-trained language models and advanced contextual embedding techniques. An ensemble model initially designed by the authors (Herath et al., 2020) is expanded here to identify offensive language. The final ensemble model contains three sub-models: the Outer Model (Classifies a post as Bullying or Defending), the Bullying Model (Classifies a post as 'Harasser' or 'Bystander assistant'), and the Defending Model (Classifies a post as 'Victim' or 'Bystander defender'). The 'defending model' demonstrates promising performance, while the 'bullying model' struggles to classify bystander assistants effectively.

When it comes to detecting cyberbullying, [21] uses Facebook comments as input. After preprocessing, feature extraction is performed to identify elements like pronouns, adjectives, nouns, and shorthand text. The presence of bullying words in preprocessed comments is tested using the LSA technique in natural language processing. Shorthand text and emoticons are then detected and classified into categories like Flaming, Denigration, Stalking, and Trickery using a random forest algorithm. The goal is to encourage bystanders to prevent further victimisation and the spread of harmful gossip on social media. While the aim is mentioned, there's no specific categorisation done for Bystander Detection.

In [22], a Dutch and English corpus is constructed through crawling. A fine-grained annotation scheme [19] with two levels of annotation is performed. Feature types like Word n-gram bag-of-words, Character n-gram bag-of-words, Term lists, Subjectivity lexicon, and Topic model features are extracted after preprocessing the corpus. Binary classification experiments use a linear kernel support vector machine (SVM) for automatic cyberbullying detection. Two ensemble learning techniques, Voting and Cascading Classifiers, are also investigated. Multiclass classification algorithms, including a linear-kernel SVM, an LR, a passive-aggressive (PA), and a stochastic gradient descent (SGD) classifier, are tested. Text classification experiments are performed for English using pre-trained BERT, RoBERTa, and XLNet models. For Dutch, BERTje and RobBERT models are tested.

In [23], the authors explore the role of group dynamics in influencing the toxicity of Twitter conversations. They examine how bystanders and the tone of early comments contribute to spreading toxic behaviour on Twitter. The significance of social norms in predicting online human behaviour and how users respond to uncivil comments or abusive language is emphasised. They hypothesise that the number of users participating in a conversation before encountering the first toxic reply impacts whether users feel compelled to respond to a toxic reply. The hypotheses include the number of users participating before observing the first toxic reply negatively affecting the number of users posting non-toxic replies afterwards (H1), posting a non-toxic reply immediately after a toxic reply leading to more non-toxic replies (H2), and posting a non-toxic reply after a toxic reply making the conversation more likely to become non-toxic (H3). The findings suggest a bystander effect, where a higher number of participants before a toxic tweet is associated with fewer users responding toxically. They find that early reactions to toxic tweets within conversations are crucial. The posting of a toxic response immediately after a toxic comment is connected with users posting non-toxic replies and conversations becoming increasingly toxic. The study indicates that posting a toxic reply after a toxic comment predicts that the Twitter conversation will become more toxic. This underscores the potential for people to ignore or respond sympathetically to incivility, potentially preventing escalating arguments.

Challenges in Bystander Detection

We can observe from the previous studies that machine-learning approaches have not been applied to classify bystander roles; only machine-learning techniques have been used for cyberbullying detection. This presents a notable research gap. The majority of previous studies have utilised annotation techniques to classify bystanders and their types. The availability of data also influences the strategies employed for classifying bystanders. The existing data is not filtered and up-to-date, making it challenging to categorise individuals who do not respond to content on social media. Constructive interactions from these unaware bystanders have the potential to significantly mitigate the effects of cyberbullying on victims. Recognising that no one-size-fits-all mechanism can be applied across all social networking sites is important.

Table 1: Datasets for Cyberbullying Research

Data Source	Data size	Data Language	Data Gathering Tools
ASKfm [19]	91,370 Dutch posts	Dutch	GNU Wget software
ASKfm [20]	-	English	AMICA
Facebook [21]	100 comments	English	-
ASKfm [5, 22]	113,698 English posts and 78,387 Dutch posts	English and Dutch	GNU Wget software
Twitter [23]	79,799 conversations with 528,041 tweets	English	Twarc

Table 2: Cyberbullying detection and bullying classification research work

Characteristics	Preprocessing steps	Classifier	Methods	Cyberbullying Types
Bag-of-word & polarity based on sentiment lexicon features are extracted to detect cyberbullying & binary classifiers were built for each of the categories [19]	Tokenization PoS-tagging and lemmatisation	Binary Classifier	SVM	Harasser, victim, and Bystander
The prebuilt ensemble model is extended with a pre-trained BERT embedding layer, a hidden neural layer and a softmax output layer [20]	Replacing slang words and abbreviations, decoding emoticons, Removal of punctuations, Upper to lower case, tokenisation and special token additions	Binary Classifier	Ensemble model	Harasser, Victim, Bystander, defender and Bystander assistant

Cyberbullying is detected in social media platforms by utilising Latent Semantic Analysis, multitask multimodality Gated Recurrent Unit and Dirichlet Multinomial Mixture, which assists many end-users in avoiding becoming a victim of cyberbullying [21].	Tokenisation, lemmatisation, removing special characters, stop words and stemming.	Random Forest	Latent semantic and analysis feature extraction	Denigration, Trickery, Flaming and Cyberstalking
Automatic cyberbullying detection in social media text by modelling posts written by bullies, victims, and bystanders [5].	Tokenisation, PoS-tagging, lemmatisation, removal of hyperlinks, white spaces, replacement of abbreviations with full form, sentiment lexicon matching, and stemming	Binary Classifier	SVM	Harasser, Victim, Bystander, defender and Bystander assistant
A bystander effect was discovered, demonstrating a negative correlation between the number of Twitter users who participated in the conversation before a toxic tweet was sent and the number of people who responded to the toxic tweet in a non-toxic manner. Additionally, it was discovered that how people react in the first instance to harmful tweets matters a lot [23].	Removed tweets with links, images, and videos instead of text.		Multivariate regression analysis, Poisson regression model and linear regression model	Bystanders and Proposed Three Hypothesis
A series of multiclass classification experiments to determine the feasibility of text-based cyberbullying. Participant role detection. The performance of feature-engineered single and ensemble classifier setups and transformer-based pre-trained language models (PLMs) are investigated [22].	Tokenization, part-of-speech-tagging, and lemmatization	Linear classification, Voting classifier and Cascading classifier	SVM, Logistic regression, passive-aggressive, SGD BL and Random Majority BL	Harasser, Victim, Bystander defender and Bystander assistant

Table 3: Evaluation metrics analysis

Ref	Accuracy	F1-score	Precision	Recall
[19]	78.50%	55.39% for cyberbullying and 35.09% for the "defence" category	60% for cyberbullying	51% for cyberbullying

[20]	-	83% for cyberbullying, 76% for role classification	84% for cyberbullying, 68% for bystander-defender	82% for cyberbullying, 73% for bystander-defender
[21]	70%	70% for cyberbullying	-	-
[5]	57.19% for English, 94.47% of Dutch	64% for English, 61% for Dutch	73.32% for English, 56.76% for Dutch	57.19% for English, 66.40% for Dutch
[22]	-	55.19% for English, 53.31% for Dutch	59.90% for English, 59.18% for Dutch	56.80% for English, 50.48% for Dutch

Conclusion and Future Work

In this paper, a comprehensive literature overview study was conducted on cyberbullying and its detection using machine learning techniques. The severe impacts of cyberbullying are evident as numerous teenagers and adolescents become victims of bullying for various reasons, leading to grave consequences such as suicides, depression, and mental disturbances. The study revealed that only a limited number of research studies have focused on the role classification of cyberbullying, including categorising bystanders. The availability of suitable data also poses a challenge.

Several key points emerge from the overview that warrant emphasis in future research. Firstly, there is a need to establish a publicly accessible dataset specifically curated for cyberbullying research. This dataset should encompass content from a diverse array of social media platforms. Secondly, the development of advanced machine learning methods for the detection and prevention of cyberbullying is crucial. This includes the incorporation of role classification for the author of the posts. Lastly, there is a call for applying Machine Learning techniques, rather than Annotation Techniques, to detect Bystander Assistants and Bystander Defenders. This shift can enhance the accuracy and efficiency of such detection methods.

References

[1]. Mahlangu, T., Tu, C., & Owolawi, P. A. (2018). A review of Automated Detection Methods for Cyberbullying. 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC).

[2]. What is cyberbullying (2021) StopBullying.gov. Available at: https://www.stopbullying.gov /cyberbullying/what-is-i t (Accessed: 16 July 2023).

[3]. Bystanders are essential to bullying prevention and intervention (2021), StopBullying.gov. Available at: https://www.stopbullying.gov/resources/research -resources /bystanders-are-essential (Accessed: 16 July 2023).

[4]. Obermaier, M., Fawzi, N., & Koch, T. (2016). Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying. New Media & Society, 18(8), 1491–1507.

[5]. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic cyberbullying detection in a social media text, PLOS ONE, 13(10),e0203794.

[6]. Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry, 49(4), 376–385.

[7]. Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. Archives of Suicide Research,14(3),206–221.

[8]. [8] Kopecký, K., & Szotkowski, R. (2017). Cyberbullying, cyber aggression and their impact on the victim – The teacher. Telematics and Informatics, 34(2), 506–517.

[9]. Nadali, S., Murad, M. A. A., Sharef, N. M., Mustapha, A., & Shojaee, S. (2013). A review of cyberbullying detection: An overview. 2013 13th International Conference on Intelligent Systems Design and Applications.

[10]. Haidar, B., Chamoun, M., & Yamout, F. (2016). Cyberbullying Detection: A Survey on Multilingual Techniques. 2016 European Modelling Symposium (EMS).

[11]. Different types of cyberbullying End cyberbullying. Available at: https://endcyberbullying.org/5-different-types-of-cyberbullying/ (Accessed: 18 July 2023).

[12]. Zainudin, N. M., Zainal, K. H., Hasbullah, N. A., Wahab, N. A., & Ramli, S. (2016, May 1). A review of cyberbullying in Malaysia from a digital forensic perspective. IEEE Xplore.

[13]. Walisa Romsaiyud, Kodchakorn Na Nakornphanom, Pimpaka Prasertsilp, Piyaporn Nurarak, & Pirom Konglerd. (2017). Automated cyberbullying detection using clustering appearance patterns.

[14]. Willard, N. E. (2007). Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress. Research Press.

[15]. Kansara, K.B., & Shekokar, N.M. (2015). A Framework for Cyberbullying Detection in Social Networks.

[16]. Goebert, D., Else, I. R. N., Matsu, C., Chung-Do, J. J., & Chang, J. Y. (2010). The impact of cyberbullying on substance use and mental health in a multiethnic sample. Maternal and Child Health Journal, 15(8), 1282–1286.

[17]. Van Hee, C., Verhoeven, B., Lefever, E., De Pauw, G., Daelemans, W., & Hoste, V. (2015). Guidelines for the Fine-Grained Analysis of Cyberbullying. https://www.lt3.ugent.be /media/uploads/publications/2015/Guidelines_Cyberbullying_TechnicalReport_1.pdf.

[18]. Salmivalli, Christina, et al. "Bullying as a Group Process: Participant Roles and Their Relations to Social Status within the Group." Aggressive Behavior, vol. 22, no. 1, 6 Dec. 1996, pp. 1–15.

[19]. Van Hee, C., Lefever, E., Verhoeven, B., & Hoste, V. (2015). Automatic detection and prevention of cyberbullying. ResearchGate. https://www.researchgate.net/publication/320922834_ Automatic_detection_and_prevention_of_cyberbullying.

[20]. Ratnayaka, G., Atapattu, T., Herath, M., Zhang, G., & Falkner, K. (2020). Enhancing the Identification of Cyberbullying through Participant Roles. arXiv (Cornell University).

[21]. Identification and Classification of Cyberbully Incidents Using Bystander Intervention Model. (2019). International Journal of Recent Technology and Engineering, 8(2S4), 1–6.

[22]. Jacobs, G., Van Hee, C., & Hoste, V. (2020). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? Natural Language Engineering, 28(2), 141–166.

[23]. Aleksandric, A., Singhal, M., Groggel, A., & Nilizadeh, S. (2022). Understanding the bystander effect on toxic Twitter conversations. arXiv (Cornell University).

Biographies

MS KHUSHBOO THAKKAR received a BE degree in Computer Science and Engineering in 2017 from Government Engineering College, Patan, affiliated to Gujarat Technological University, Ahmedabad and received an ME degree in Cyber Security from Graduate School of Engineering and Technology, affiliated to Gujarat Technological University, Ahmedabad, in 2022. Currently, she is working as a Resource Person in PG Diploma in Cybersecurity and Law, SSCBS, University of Delhi, Delhi. She has published two books on Digital Forensics and its practical approach. Her area of interest is Cyber Security and Digital Forensics. Khushboo Thakkar may be reached at khushboo@sscbs.du.ac.in

DR ANAMIKA GUPTA received an MCA degree from the University of Delhi, India, in 1996 PhD in Data Mining from the University of Delhi, India, in 2013, respectively. She is an Associate Professor of Computer Science at SSCBS, University of Delhi. Her teaching and research areas include machine learning, image processing, and Data Analysis. She has co-authored NCERT computer science textbooks and published over 20 book chapters, journal and conference papers. Dr Anamika Gupta may be reached at anamikargupta@sscbsdu.ac.in

MR VIBHOR MATHUR, a Department of Computer Science student at SSCBS College, University of Delhi, Delhi. Vibhor Mathur may be reached at vibhor.21564@sscbs.du.ac.in

MR AMAN TIWARI, a student in the Department of Computer Science, SSCBS College, University of Delhi, Delhi. Aman Tiwari may be reached at aman.21505@sscbs.du.ac.in