International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

A Systematic Literature Review on Cyberbullying in Social Media: Taxonomy, Detection Approaches, Datasets, and Future Research Directions

Sahana V[1], Anilkumar K M[2]
[1]Department of Information Science and Engineering
JSS Academy of Technical Education
Bangalore, India
e-mail: sahana26@gmail.com
[2]Department of Computer Science and Engineering
JSS Science and Technology University
Mysuru, India
e-mail: anilkm@sjce.ac.in

Abstract— In the area of Natural Language Processing, sentiment analysis, also called opinion mining, aims to extract human thoughts, beliefs, and perceptions from unstructured texts. In the light of social media's rapid growth and the influx of individual comments, reviews and feedback, it has evolved as an attractive, challenging research area. It is one of the most common problems in social media to find toxic textual content. Anonymity and concealment of identity are common on the Internet for people coming from a wide range of diversity of cultures and beliefs. Having freedom of speech, anonymity, and inadequate social media regulations make cyber toxic environment and cyberbullying significant issues, which require a system of automatic detection and prevention. As far as this is concerned, diverse research is taking place based on different approaches and languages, but a comprehensive analysis to examine them from all angles is lacking. This systematic literature review is therefore conducted with the aim of surveying the research and studies done to date on classification of cyberbullying based in textual modality by the research community. It states the definition, , taxonomy, properties, outcome of cyberbullying, roles in cyberbullying along with other forms of bullying and different offensive behavior in social media. This article also shows the latest popular benchmark datasets on cyberbullying, along with their number of classes (Binary/Multiple), reviewing the state-of-the-art methods to detect cyberbullying and abusive content on social media and discuss the factors that drive offenders to indulge in offensive activity, preventive actions to avoid online toxicity, and various cyber laws in different countries. Finally, we identify and discuss the challenges, solutions, additionally future research directions that serve as a reference to overcome cyberbullying in social media.

Keywords-Cyberbullying; Sentiment analysis; Social media; Literature review; Machine learning.

I. INTRODUCTION

The area of Natural Language Processing (NLP) called Sentiment Analysis (SA) analyzes and studies the emotions, attitudes, appraisals, and assessments that individuals express in writing [1]. The use of social media has substantially increased due to internet growth. In social media platforms like Twitter, Facebook, and Instagram, millions of people prefer to express their opinions online and interact socially on day-to-day basis, which has increased the quantity of online social interactions and communications. Internet users' lives are becoming increasingly influenced by social media. Figure 1 illustrates the timeline of the global internet population [2]. The number of social media users globally is estimated to reach 5.85 billion by 2027, based on the most recent data [3], as shown in Figure 2. On the basis of the number of global active users (in millions) [4], figure 3 shows the ranking of social media platforms. Also, social media platforms offer a way for ideas and thoughts that would otherwise go unspoken and neglected by traditional media to be heard and explored [5]. In spite of the majority of positive outcomes that social media and the internet have yielded to society, there have been some negative outcomes as well. It is becoming more common to read interaction and communication content that indicates upsetting, disturbing, and negative phenomena such as online cyber-hate, harassment, cyberbullying, stalking, and cyber threats [6]. As a result, various categories of users have been attacked based on their religion, ethnicity, social status, age, etc. As a result of such offenses, individuals often struggle to cope with their consequences. In order to detect online cyberbullying and cyberhate speech, several NLP-based approaches have been implemented. This is because dealing with and eliminating unpleasant communications might be made easier and more convenient by using computational linguistic analysis to quickly identify and classify offenses. [7].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
Article Received: *25 July 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *06 October 2023*

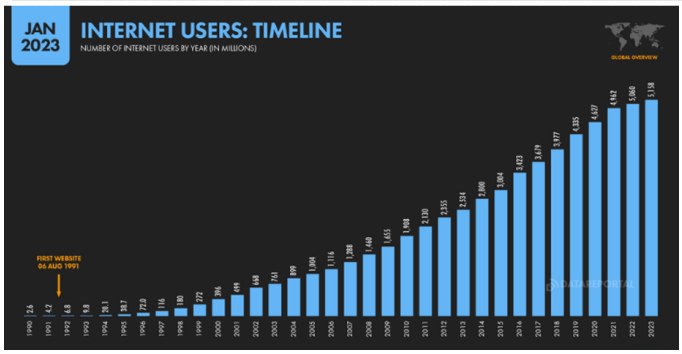

Figure 1. Number of Global Internet Users by Year (in Millions)

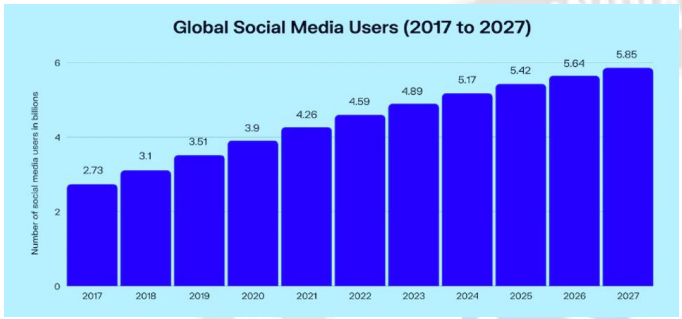

Figure 2. Number of Global Social Media Users Within 2017–2027

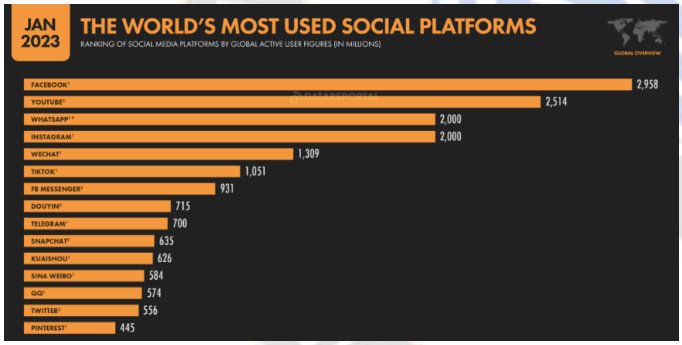

Figure 3. The world's most used Social Media Platforms

One of the online offensive behaviours is cyberbullying, which is described as the usage of electronic communications tools to hurt or victimise someone. It can attack entire communities and also specific people [8]. When someone uses the internet to harass or upset a child or young person, it is called cyberbullying. It can happen on any online or electronic service or platform, including social media sites, games, apps, and other digital content. Posts, comments, texts, messages, conversations, livestreams, memes, pictures, and videos are few examples. Listed below are some instances of how the internet has been used to impair someone's self-confidence:

- Expressing disparaging remarks about them.
- Dishonoring them by posting pictures or videos.
- Using the internet to spread false information about someone.
- Setting up phoney accounts in their name.
- Tricking people into thinking they are someone else.

Through cyber innovation and technology, several kinds of violations are committed for a variety of causes in the online cyber world. In order to reduce the pain and damage that these crimes cause to society, it is important to take into account the unsafe environment of online social networks. To get the best outcomes for automated cyberbullying detection using machine learning approaches, several researchers are researching in different directions. A taxonomy of various methods for English-language cyberbullying detection will be described in this publication.

The contributions of the paper are as follows:

- We briefed about the various types of offensive behaviour on social media platforms.
- Definitions, Taxonomy and roles/variables related to cyberbullying.
- The research foci in cyberbullying research.
- Categorised broadly the methods for detecting cyberbullying from our perspective and reviewed recent research papers that fall into these categories.
- Cyberbullying classification was largely based on which dataset?
- What was the dataset's size?
- Which classification method was used? And How often has each been used?
- Measures of quality that were used.
- Effectiveness of each approach.
- We compiled the motivating elements that lead offenders to act in an offensive manner.
- Discussions include some preventative measures to improve the environment on online social networks and cyber laws from different nations that enforce severe punishments.
- Finally, we identified the problems that still need to be solved in terms of identifying offensive online behaviour on social networking sites. In the future, it would be beneficial for researchers to tackle these issues.

The rest of the manuscript is organized as follows: Section 2 provides a summary of the types of offensive behavior in social media, a detailed taxonomy of cyberbullying, analysis of cyberbullying in comparison to other forms of bullying and a comprehensive framework of factors/variables that may influence cyberbullying. The literature search and identification processes are then described in Section 3. Section 4 discusses the state of cyberbullying research, which previews the research foci of the existing literature, available datasets of cyberbullying in English language, the main approaches of cyberbullying

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

detection/classification, comparative analysis of the literature review and taxonomy of cyberbullying detection techniques. The causes of persons engaging in offensive online activity, along with the preventive strategies and legal frameworks used by different nations to address it, are all covered in Section 5. Open Challenges and limitations in cyberbullying detection are discussed in Section 6. Section 7 points out the tasks that still need to be addressed. Section 8 concludes this manuscript.

II. DEFINITIONS AND CONCEPTS

A. Types of offensive behavior in social media

The following are some prevalent examples of offensive behaviours that people who use social media frequently engage in. Offensive behaviour can take many different forms [9]:

(1) *Abuse:* When technology is used to harm or distress someone, the behaviour is referred to as abuse and is known as online abuse. Anywhere, including social media, messaging services, email, gaming apps, live streaming websites, etc., it can happen. However, abuse primarily affects young people who are in a relationship.

(2) *Hate Speech:* It is a form of internet communication intended to disparage a group or an individual according to their gender, ethnicity, religion, race, sexual orientation, or other characteristics [10].

(3) *Misogyny:* It is showing hatred for women. To maintain women's inferior status to that of men is sexist [11].

(4) *Xenophobia:* It is a manifestation of irrational animosity towards outsiders [12].

(5) *Troll:* It is an unsettling act of reaction to something posted online. Simply put, it tends to stir things up. Fake accounts are frequently created instantly solely for this purpose [13].

(6) *Cyber aggression:* Online bullying between classmates generally occurs sporadically. There won't be an imbalance of power between the aggressor and the target, and there won't be any malice or desire to cause trouble [14].

(7) *Cyber bullying:* With the intention to hurt the target through digital devices, it is a malicious and targeted method [15]. It involves using technology to email, post, or share offensive, defamatory, cruel, hurtful, embarrassing, or threatening audio, video, or textual content [16,17]. Any online platform, including social media, forums, apps, etc., can experience it. One of the genuine cases of cyberbullying involves a 13-year-old girl named Megan who, from the age of 8, had experienced significant depression. To tease Megan, a female neighbour set up a boy's name MySpace account. However, Megan finds herself drawn to the account holder she initially mistook for a boy. However, that person then turned-on Megan and began speculating about rumours. Even though it all started out as a private conversation, over time, that person began releasing private conversations publicly and said “the world would be a better place without her”. After telling the guy he was the kind of boy a girl would kill herself over, Megan committed suicide over that remark [18].

B. Definitions and roles in cyberbullying

In cyberspace, bullying has been described in many different ways: Electronic bullying [19], Internet bullying [20], Internet harassment [21], online bullying [22], and online social cruelty [23], with cyberbullying being the most popular among researchers. The terminology used in most studies on cyberbullying was taken from the literature on traditional bullying [25,26]. Based on the study [27], Social Networking Sites (SNSs) bullying is a type of cyberbullying that takes place on SNSs. It is characterized as deliberate, hostile behaviour on group of individuals or an individual which involves repeatedly sending aggressive content meant to hurt or discomfort a target through social media. A social media cyberbullying case is shown in figure 4. Cyberbullying frequently comprises reciprocal relationships between perpetrators, victims, and bystanders as a triadic system due to the interconnection of SNSs and their capacity to involve several people in social interactions over the internet [28, 29]. A perpetrator is an individual or a group of people who purposefully cause another person pain or anguish on a regular basis. A bystander is someone who witnesses bullying and has the choice to either (1) participate in the bullying, (2) Encourage the perpetrators or consolate the victims, or (3) disregard the incident. Victims are people who repeatedly receive hurtful and power-imbalanced messages and experiences.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

Figure. 4. A real tweet from the social media platform Twitter that contains an actual instance of cyberbullying. To conceal the identities of the individuals engaged in the cyberbullying case, some portions of the photograph have been blurred [30].

C. Taxonomy of cyberbullying

There are ten different varieties of cyberbullying, as indicated in Figure 5, that vary from gossiping about or excluding somebody to making fun of their race or religion.

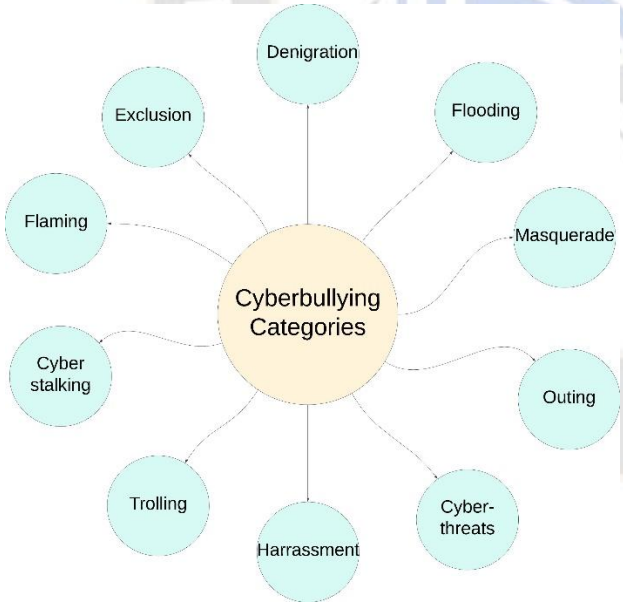

Figure 5. Cyberbullying Categories

Below is a list and definition of the categories that make up the taxonomy of the phrase "cyberbullying":

i. Exclusion is when the victim is ignored or neglected in a conversation [31]. Cyber-Exclusion is the deliberate and intentional use of technology to inform individuals that they are not a part of the group and that their participation is not required. On social networking sites, users have the option to unfriend or block others, which prevents them from viewing their profiles or leaving comments.

ii. Denigration is the act of insulting, slandering, disparaging, or degrading another person on social media [32]. Denigration is when someone is disparaged by writing nasty, vulgar, hateful, cruel, or false remarks or rumours about them to someone else or by posting them in a chat room or public forum. Since the insults are visible to others in addition to the victim, the goal is to harm the victim's reputation in the eyes of his or her community.

iii. Flooding is the act of sending a huge volume of social networking messages at once to prevent the victim from submitting anything [33]. The enter key is held down while the bully or harasser writes the same comment over and over again, posts irrelevant comments, or otherwise prevents the victim from participating in the chat or conversation.

iv. Masquerade is the act of pretending to be someone else in order to convey messages that damage or create harm and appear to have come from that person [34]. One way to accomplish this is, for instance, to hack into the victim's email account and send these messages right away. This form of access can also frequently be achieved by friends swapping passwords; however, a skilled hacker may find other methods, such as by repeatedly testing potential passwords. This tactic is inherently challenging and challenging to detect or identify.

v. Flaming, blazing, or battering involve at least two users physically and verbally assaulting one another. Flaming is a form of cyber-hate that includes communications and interactions that are frequently rude personal attacks and are aggressive, unpleasant, indignant, angry, and insulting [35]. Flaming can happen in a variety of settings, including email, Twitter, discussion forums, and online social networking sites. Capital letters are frequently used to convey anger, as in the phrase "U R AN IDIOT & I HATE U!" Numerous flame messages are violent, abhorrent, and cruel and lack any basis in reality or logic.

vi. Cyberstalking is the practice of following, bothering, or harassing someone online through the use of social media sites [36]. It could include erroneous charges, defamatory statements, criticism, slander, and libel. Cyberstalking incidents sometimes start as seemingly innocent contacts. Occasionally, especially in the beginning, a few odd or maybe offensive remarks

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

might even be amusing. However, if they prove to be systematic, it becomes grating, infuriating, and downright frightful.

vii. Trolling, also known as baiting, aims to start a fight by purposefully posting comments that disagree with other contributions in the discussion [37]. The commenter intends to stir up feelings and a debate, but the remarks themselves invariably become intimate, crude, enthusiastic, or emotional.

viii. Denigration and outing are similar, but outing calls for a close, in-person or online, personal interaction between the bully and the victim. It involves publishing literature that contains intimate, embarrassing, or degrading information about oneself [38]. The victim's accounts or any other personal information, such as addresses, phone numbers, or passwords, may be included in this material.

ix. Social media harassment is the same as harassment committed through more traditional and conventional methods [39]. Threatening behaviour that is motivated by a person's age, gender, race, sexual orientation, or other characteristics is referred to as harassment.

x. Cyber threats are short communications that contain threats of harm, are ominous, threatening, extremely aggressive, or contain extortion [40]. The line separating harassment from cyberstalking is hazy, but one sign may be when the victim begins to fear for his or her safety or well-being. At that point, the behaviour must be classified as cyberstalking.

D. Comparing cyberbullying with other types of bullying

Compared to face-to-face bullying and various digital communication channels like phone, e-mail, and text messaging services, social media networks constitute a coherent and new communication setting [41]. SNSs features including relational linkages, search, privacy, and network transparency [42] have increased the opportunity for victims, offenders and bystanders to engage in conversation on these platforms. For example, digital profiles give offenders a higher level of anonymity, so escalating the power disparity among victims and perpetrators. Additionally, the networked platform allows other users to access bullying posts, which causes victims to suffer harm repeatedly. Through qualities that are rare in digital communication medium, such as sharing, liking and hash-tagging, SNSs also allow for the rapid dissemination of bullying content to a larger audience. As a result of these behaviors, bullying is no longer understood as a dyadic relationship based on face-to-face communications or other digital communication media contacts, but as a triadic reciprocal relationship based on communication media [43].

The following subsections compare cyberbullying to other forms of bullying and discuss how SNSs aggravate its detrimental effects from six aspects: intentionality, repetition, power imbalance, anonymity, accessibility, and publicity [44,45,28,46,47].

1. Intentionality

Intentionality is defined as the desire to cause another person harm [48]. In face-to-face bullying, the bully's intent of harming the victim is obvious, as in beating her/him up, however in bullying through digital communication media, the bully's intent to hurt is triggered when the victim receives and reads the bully's words. Such harm-intention is overt and unambiguous. On social networking sites, users often update their digital profiles by posting articles related to their interests or uploading images of themselves. There are more potential for people to encounter SNS bullying than face-to-face and digital communication media bullying due to the regular and persuasive usage of SNSs that results in a constant revelation of personal information. For example, an SNS post can be hurtful if a perpetrator makes derogatory comments about a user's physical appearance. The perpetrator's behaviour (the "wants to hurt" part) is where the intentionality begins. A teasing meme that makes reference to someone's appearance and spreads through their social network can also be upsetting [42]. Thus, the victim's perspective ("felt hurt") is used to interpret the purpose to harm [42,49].

2. Repetition

Repetition describes intentional, repeated behaviours that harm a target. When someone physically hurts or hits a person on several occasions, it is considered repetition in face-to-face bullying. In SNS bullying or digital communication media, repetition occurs when harmful content is repeatedly passed on or spread [50,51]. The repetition of bullying content on SNSs is possible due to the fact that users are able to share, view, or respond to the content. In this way, SNS may offer a greater opportunity for bullying than face-to-face communication or digital communication. Redistributing humiliating things on social networks where the bystanders and perpetrators are connected can be used to achieve repetition [42]. In addition, the content can be shared and read on social networks numerous times, resulting in the bullying behaviour repeating itself.

3. Power imbalance

The term power imbalance refers to a situation in which a powerful person bullies a less powerful individual [52]. Two perspectives can be considered when analyzing a power imbalance: (1) the victim's insufficient power in comparison to the perpetrator's; and (2) the perpetrator's greater power [48]. Bullying, in whatever form it may take, usually involves perpetrators who are stronger in relational, social and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

psychological terms [53]. A disparity in power means that those who engage in bullying on digital communication media and social networking sites may also be highly skilled in technology [54]. In SNS bullying, power imbalances can result from the social network's features, such as a digital profile, which enables perpetrators to separate their online and offline identities. For example, virtual private network (VPN) can be used for hiding one's location and manipulating one's offline-online identity. Furthermore, SNSs facilitate the continuous dissemination of bullying content, enabling easy access to it [42]. Due to the difficulty in stopping the endless cycle of bullying on a proactive basis, SNS bullying victims [55] feel powerless to stop it [56].

4. Anonymity

Anonymity describes the extent to which one's identity is unknown. Due to victims' ability to recognize the perpetrator's voice, appearance and stature, face-to-face bullying is harder for perpetrators to conceal their identities. A network service provider can identify a person who commits digital communication media bullying. The perpetrator of cyberbullying, by contrast, can remain relatively anonymous. For example, perpetrators can conceal their identity by posting comments and sharing images anonymously on SNSs like 4chan, which is essentially an image-based social network. Perpetrators can also use pseudonyms on SNSs. Even though popular SNSs like Facebook only allow users to create their own profiles with their legal names and photos, users can still remain anonymous by creating a different account by using fictitious identity documents. As it is easy to keep one's real identity and legal names separate from their online personas on SNSs, abusers can hurt victims without worrying about being held accountable.

5. Accessibility

Accessibility is easy access to a target. Victims of face-to-face bullying can avoid the abuse by finding a safe haven [57]. Digital communication victims can use a new email address or phone number to stop receiving harassing calls or emails. Social networking bullying, however, is neither confined to physical spaces nor to routine, common interactions [44,58]. Due to their limitless connectivity, social networking sites provide bullies with an opportunity to bully anyone, anytime, anywhere, with or without victims present [54,59]. Although victims can deactivate their accounts permanently, bullying content can remain on the platform and be shared by perpetrators. It is hard for victims to escape humiliation on social networks due to "users' ability to view and traverse their connections and those made by others on the platform" [42].

6. Publicity

The term "publicity" describes how many people were made aware of a bullying incident. An incident of face-to-face bullying may only involve other students in the same class or only involve coworkers who are employed by the same organization. Bullying through digital communication means that the victim is the only one who can hear or read humiliating calls, texts, and emails. Public broadcasts of bullying-related content are unlikely. SNS gives a bully multiple way to publicize their bullying behaviour. For example, a perpetrator could post edited pictures of a victim and ask SNS users to check them out and comment. The pool of contacts of other users that the victim, perpetrators and bystanders have in common on the platform allows harassing messages to reach an infinite number of people with hashtags and tags [42].

E. A comprehensive framework of variables that influence cyberbullying

Triadic reciprocal connections between cyberbullying perpetrators, victims, and bystanders on online platform make the phenomenon complicated. With the help of Social Cognitive Theory (SCT) [60], a comprehensive framework for explaining key constructs has been developed.

"A conceptual framework within which to analyze the determinants and psychosocial mechanisms through which symbolic communication influences human thought, affect, and action" is provided by SCT [60]. Cyberbullying involves three types of participants, as depicted in Figure 6: perpetrators, victims, and bystanders. Perpetrators are individuals who intentionally harm victims who are unable to fortify themselves on a regular basis. Victims are people who consistently encounter hurtful interactions with perpetrators and bystanders. Bystanders are third parties who observe cyberbullying and has the choice to take action that could affect how the incident develops (for example, challenging the bullies or offering support to the victims, participating in the bullying or choosing to ignore the bullying incident).

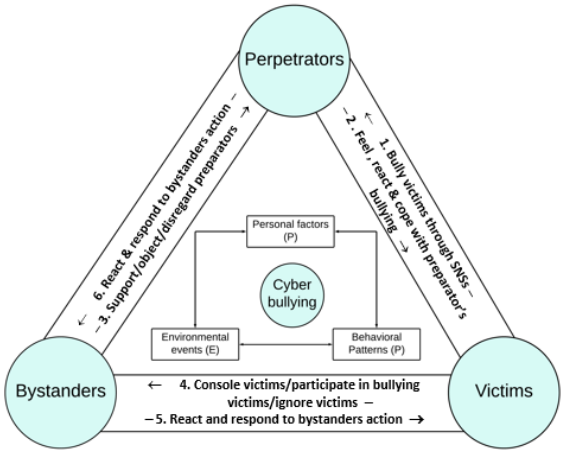

Figure. 6. Triadic reciprocal connections among perpetrators, victims, and bystanders

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

According to SCT, personal factors (P), behavioural patterns (B) and environmental events (E) all have a triadic reciprocal influence on one another and shape how people behave. Alternatively, "people are producers as well as products of social systems" [60]. Cyberbullying personal factors (P) include socio-demographic characteristics, beliefs, expectations, goals, self-perceptions, thoughts and emotions that a bully, victim, or bystander brings to the situation. An environmental event (E) can refer to a wide range of incident-related and situational cues characterized by bullying, including participant traits, technological inputs and social cues. Behaviour patterns (B) are the reactions and behaviours that the bully, the victim, and any bystander exhibit before, during, and after cyberbullying.

A cyberbullying incident often begins with a perpetrator posting offensive content on social media sites, like humiliating messages or embarrassing photos, as seen in Route 1. Victims are prone to the bullying content of the perpetrators, which affects how they feel, act, react, and cope, as in Route 2. The content of cyberbullying, as well as the characteristics and responses of the victims, is exposed to others in SNSs (i.e., bystanders) who are linked to the perpetrators and victims. Based on this, they may choose to console the victims, disregard the incident, or participate in the perpetrator's bullying behavior as a result, as in Routes 3 and 4. The victims and perpetrators on SNSs eventually identify the bystander's behaviours, which in turn influences their ideas, emotions, and behaviours in a triadic reciprocal manner, as in Routes 5 and 6.

- It incorporates key elements of theories employed to research perpetrators (such as social learning model of deviance [61] and the crime opportunity theory [44]), victims (such as the transactional theory of stress and coping [62]), and bystanders (such as the bystander effect [63,64] and just world belief [65]). Consequently, it is a comprehensive framework that aids in pragmatic consolidation of the broad range of variables pointed out in the literature.
- Many types of cyberbullying and bullying behaviors have been cited to explain it in a variety of contexts [66–72]. The factors related to cyberbullying perpetrators, victims, and bystanders can be thoroughly examined and described as well as research gaps and opportunities can be identified for future research.

F. Outcomes of cyberbullying

Cyberbullying can result in serious detrimental effects on both the bully and the victim's mental health. Because it can quickly escalate into indirect cyberbullying and it is impossible for the victim to flee, the negative impacts of cyberbullying are more detrimental than those of traditional bullying. The main negative impacts of cyberbullying are observed to be depression and isolation. Numerous victims, according to D. Mann, report having emotional, behavioural, and concentration problems [73]. These victims have probably also complained of regular headaches, ongoing stomach pain, and trouble sleeping. According to the author Tjhin Wiguna and coworkers survey, the most severe effect of cyberbullying may involve a suicidal attempt. [74]. The impact of cyberbullying, according to authors, is that male victims and perpetrators become more aggressive and develop alcohol or cigarette addictions, whereas female victims exhibit internalising behaviours like ideation, isolation, depression, or suicidal thoughts. Higher levels of cyberbullying have been associated with elevated levels of depression, according to research [75]. Additionally, according to author Nixon, 32% of targets of cyberbullying displayed at least one stress symptom.

On social media platforms, teenagers and adults are equally vulnerable to cyberbullying. Nearly 50% of American youth report experiencing bullying [76]. But most of these victims typically conceal their victimisation for a variety of reasons. Teenagers and adolescents worry that they'll lose access to their device. Adults are ashamed to admit they are being bullied, and they worry that their peers or family members won't understand them. Therefore, it is crucial to recognise and report cyberbullying to track down victims and intervene to treat them. Additionally, it is necessary to protect society from the harm that cyberbullying is causing.

III. LITERATURE SEARCH AND IDENTIFICATION

In this section, we outlined the main stages of the literature review process, as illustrated in Figure 7.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

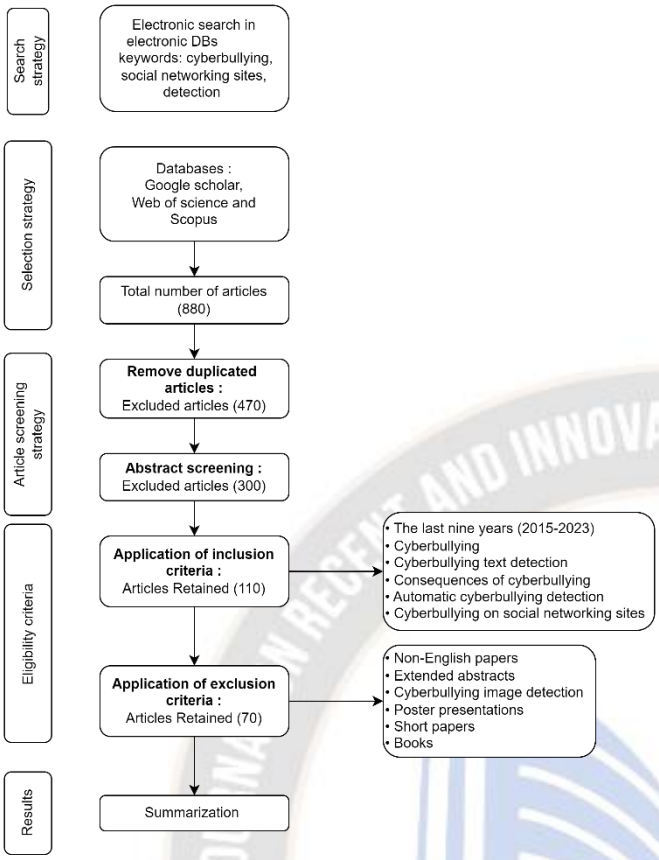

Figure. 7. The outcome of the PRISMA flow illustration

A. Search Strategy

An extensive search for research papers on the detection of cyberbullying in social networking sites was conducted. For the current study, the authors chose the Web of Science, Scopus and Google Scholar databases owing to their thoroughness and multidisciplinary coverage. In total, 880 documents were initially retrieved from both databases, as shown in Table 1.

TABLE I. SEARCH PARAMETERS

Database	Search term	Search inside	Date array	Query expression	Sort by	Number of documents
Web of Science, Scopus and Google Scholar	Cyber bullying	All fields	2015 to 2023	ALL (cyberbullying AND detection AND in AND social networking sites) AND PUBYEAR > 2015 AND PUBYEAR < 2023	Relevance	880

B. Selection strategy

The documentation required for the literature selection approach was identified using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram. The source of databases was selected in accordance with the relatedness of computer science and information technology areas. In the final phase of the selection strategy, 880 documents in total have been retrieved.

C. Article screening strategy

This stage consists of two main phases: retrieved articles phase and initial selection phase. The research papers from the aforementioned databases that were retrieved and saved were not pre-processed in any way. This systematic review began with the initial step of filtering and selecting the most relevant papers. We selected related and relevant research articles based on the abstract, which provides an overview and summary of the research. The article database has been streamlined by removing duplicates and non-relevant articles. After the article screening strategy phase, 110 articles were retained.

D. Eligibility Criteria

During the eligibility criteria stage, inclusion and exclusion criteria were developed to ascertain whether the included study was valid and relevant to our main goal.

1. Inclusion criteria

For publications to be considered for inclusion, they must satisfy the following criteria:

1) The last nine years (2015-2023)
2) Cyberbullying
3) Cyberbullying text detection
4) Cyberbullying on social networking sites
5) Consequences of cyberbullying
6) Automatic cyberbullying detection

2. Exclusion criteria

The following publications were left out of the study's publication selection:

1) Non-English papers
2) Extended abstracts
3) Cyberbullying image detection
4) Poster presentations
5) Short papers
6) Books

E. Result and summarization

A final stage entailed retrieving the full text of the paper to identify the methods, techniques, approaches, and datasets used in it. To clarify any confusion or ambiguity on identifying the research articles, all discussions among team members were conducted at this stage. Additionally, based on

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
Article Received: *25 July 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *06 October 2023*

existing approaches used by scholarly literature, an initial idea of how to categorize research articles was developed. An analysis of this categorization will be presented in the next chapter.

IV. THE STATE OF CYBERBULLYING RESEARCH

A. Research foci of the existing literature

According to the articles identified, cyberbullying can be categorized into four distinct areas of research. The research focuses of the identified studies of cyberbullying are summarized in Table 2. Initially, the research was exploratory in nature. These studies mainly focused on (1) cyberbullying conceptualization, patterns and prevalence (2) scenario-based and descriptive evaluations of the phenomenon. For example, Bellmore et al. [77] used historical data of social media to investigate the five W-questions (what, who, where, why, and when) of cyberbullying. For example, they discovered that weekday evenings were the times when cyberbullying posts were most prevalent.

Research on participant's behaviours in cyberbullying was the focus of the second line of research. In accordance with participant role approach [78], three participants—perpetrators, victims, and bystanders are present during an episode of cyberbullying. Among these studies, the majority explored the reasons behind a specific role's actions and tested various types of factors connected to perpetration, victimisation, and bystander behaviour. For instance, in their research on cyberbullying, Pabian et al. [79] examined the connections between the three dark triad personality traits—namely, narcissism, psychopathy and machiavellianism—and cyberbullying. From a victim's perspective, Camacho et al.'s [62] investigation looked at how cyberbullying victimisation affected one's satisfaction with SNSs. They discovered that the cyberbullying victim's perception of its severity significantly decreased their sense of the value and fun of using social networking sites. Bystanders' propensity to step in during an instance of cyberbullying was examined by Brody and Vangelisti [63], who found that one would be less likely to want to assist if there were many other bystanders present. The third research area examined methods for stopping and identifying online bullying. Systemic and informational aspects of preventing cyberbullying were emphasized in these studies. For example, Alhabash et al. [80] examined how anti-cyberbullying messages affected anti-cyberbullying attitudes as well as their viral reach, emotional tone and affective evaluation. Positive anti-bullying messages with high "likes" and "shares" were linked to a more anti-bullying attitude, according to their research. A cyberbullying detection algorithm was created by Balakrishnan et al. [81] using the dark triad and big five personality models. They discovered that the algorithm's detection power was increased by incorporating elements from the two models. Finally, some reviews published, specifically addressed cyberbullying. Hamm et al. [82] reviewed empirical findings on the detrimental effects of cyberbullying victimisation on one's health. The summary provided by Ioannou et al. [83] focused on the prevalence, traits, and hazards of SNS bullying.

TABLE II. AN OVERVIEW OF THE RESEARCH FOCI OF CYBERBULLYING STUDIES

Stream	Focus	Study
Exploratory (n = 15)	Prevalence, patterns, and conceptualization	[77,54,84,85,86,87,88,89,90,91]
	Descriptive/ Scenario analysis	[92,56,93,94,95]
	Perpetrators	[44,61,96,97,79,98,99,100,101,102]
Participant behavior (n = 33)	Victims	[62,99,101,102,103,104,105,106,107, 108,109]
	Bystanders	[110,111,112,59,63,64,113,114,115, 116,117,118,119,120,65]
Prevention and detection (n = 4)	Prevention	[80]
	Detection	[81,121,122]
	Well-being related effects	[82,123]
Review and synthesis (n=4)	Company policy	[124]
	Prevalence, traits, and hazards	[83]

B. Cyberbullying datasets

This section provides an overview of the datasets created over the past few years for the detection of cyberbullying using various factors, such as the size, number of classes and accessibility of the datasets, as shown in Table 3. It covers a variety of social networking sites with text-based content, including Formspring (which has a teen-focused Q&A forum), Twitter, Instagram, and Facebook, which are large microblogging platforms, WhatsApp, an instant messaging app that can run on multiple platforms, and Wikipedia talk pages, which could be characterized as a collaborative knowledge repository. Every dataset identifies a unique aspect of cyberbullying. Examples of offensive, racist, and sexist tweets can be found in Twitter datasets. Racism and sexism statuses are also present in Facebook datasets. Examples of personal attacks are present in the YouTube and Instagram datasets. Formspring datasets, however, do not specifically address a single subject.

Two subset datasets are included in the binary dataset that Mangaonkar et al. [125] proposed. 170 tweets that were bullying and 170 tweets that were non-bullying made up the first subset dataset. With 1163 non-bullying tweets and 177 bullying tweets, the second sample was unbalanced. To evaluate how the ML algorithms, perform on various dataset types, a balanced and imbalanced dataset was created. These

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

tweets were then manually labelled as "bullying" or "nonbullying" for verification purposes.

Over the course of two months, Waseem and Hovy [126] collected a dataset of tweets. They retrieved 136,052 tweets and annotated 16,914 of them, of which 3383 were sent by 613 users who were sexist, 1972 were sent by 9 users who were racist, and 11,559 were sent by 614 users who were neither sexist nor racist. They did not balance the data to present the most pragmatic dataset because hate speech is a real but rare phenomenon.

A Twitter dataset made up of tweets retrieved from the public Twitter API stream was proposed by Zhao et al. [127]. Each tweet contains one or more of the following subsequent keywords: bully, bullied, bullying. Retweets are removed by filtering out tweets that contain the abbreviation "RT". Finally, 1762 tweets are manually chosen and tagged at random from the entire Twitter archive. It is significant to remember that labelling is founded on signs of bullying. Bullying traces are defined as a response to a bullying encounter, which includes but vastly outnumbers instances of cyberbullying.

The Twitter corpus from the Content Analysis for the WEB 2.0 (CAW 2.0) dataset [129] was used by Singh et al. [128]. This corpus includes approximately 900,000 postings from 27,135 users between December 2008 and January 2009 (one XML file per user). They chose this corpus because it contains data for both textual content and social networks, also, because it has been extensively used in prior literature. The comments were left in @ format, which denotes direct paths between two people, and they randomly selected 800 files. A data set of about 13,000 messages was produced as a result. Three students were then asked to classify each message as cyberbullying twice. They assigned each post a "yes" or "no" assessing based on whether they thought it involved cyberbullying. This led to a data collection that included 4865 messages sent between 2150 user pairings.

In January and February of 2015, Al-garadi et al. [130] used Twitter to gather their data. Their data set contained 2.5 million tweets with geographic tags. In their research, they only use content that is publicly accessible and is extracted using the Twitter API in accordance with Twitter's privacy policy. Just 599 tweets were classified as cyberbullying in their dataset, while 10,007 tweets were classified as non-cyberbullying. It may be challenging for the model to classify the instances appropriately whenever there is such an uneven distribution of classes. Lack of class imbalance makes learning algorithms more probable to be overwhelmed by the major class while ignoring the minor. The normal class typically makes up the majority of the data in data sets used in real-world applications like fraud detection, instruction detection, and medical diagnosis while the abnormal class makes up the minority. A combination of oversampling the minority (abnormal) class and under sampling the majority (normal) class has been suggested as one solution to these problems.

The Instagram API and a snowball sampling method were used by Hosseinmardi et al. [131] to collect data. From a randomly selected seed node, they discovered 41 K Instagram user ids. The majority of these Instagram IDs (25 K, or 61%) belonged to users with public profiles; the remaining users all had private profiles.

Data were gathered by Zhang et al. [132] from the social networking website Formspring.me. Amazon Mechanical Turk, a website service where three workers cast one vote each to determine whether or not a document contained bullying content, gathered and classified nearly 3000 messages. As a result, the workers vote on each message in an equal number. Approximately 6.6% of the messages were classified as bullying posts by at least two employees. The messages in the original dataset were broken down into sentences by the authors, and the messages with at least one vote were given new labels. 23,243 sentences were produced as a result, with 1623 (or roughly 7%) being classified as bullying messages.

English Wikipedia was used by Wulczyn et al. [133] to create a corpus of over 100k high-quality human-labeled comments. By calculating differences over the entire revision history and extracting the new content for each revision, the data of debate comments from English Wikipedia discussion pages is generated. The collected corpus was annotated by about ten people using Crowdflower (https://www.crowdflower.com, access on 16 February 2023) into two categories: attacking and not attacking.

Using a crowd-sourced hate speech lexicon, Davidson et al. [134] compiled tweets with hate speech keywords. They divided a sample of these tweets into three groups using crowdsourcing: those that contained hate speech, those that only contained offensive language, and those that contained neither. 33,458 tweets were produced because of this dataset.

A 60 K tweet dataset of large-scale, crowdsourced abusive tweets was published by Founta et al. [135]. To effectively annotate the tweets using crowdsourcing, a better strategy is used. By utilizing such methodical techniques, the authors came to the conclusion that None, Spam, Abusive, and Hateful is the most appropriate label set in identifying abusive behaviours on Twitter, with the results showing that 11% of tweets are "Abusive," 7.5% are "Hateful," 22.5% are "Spam," and 59% are "None." 'None'/'Spam' and 'Abusive'/'Hateful' are concatenated to make this dataset ready for a binary classification task.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
Article Received: *25 July 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *06 October 2023*

The first dataset of textual hate speech to be annotated at the sentence level was presented by De Gibert et al. [136]. Sentence-level annotation makes it possible to address the smallest piece of hate speech while reducing the noise produced by other clear sentences. A total of 10,568 sentences from Storm-front were gathered and classified as hate speech or not, along with two other auxiliary types.

In order to identify English-language cyberbullying on Twitter, Banerjee et al. [137] conducted research. There are 69,874 tweets in the Twitter dataset. The selected tweets were manually labelled as "0" non-cyberbullying or "1" cyberbullying by a group of human annotators.

For the purpose of text classification, Sadiq et al. [138] presented the Cyber-Trolls dataset from Data Turks. This dataset is used to categorize tweets in order to help or stop trolls. Cyberaggressive (CA) and non-aggressive (NCA) are the two categories. There are 20,001 items in the dataset, 7822 of which are cyberaggressive, and 12,179 of which are not.

Kumar and Sachdeva [139] developed two datasets FormSpring.me and MySpace. The 13,158 messages published by 50 unique users on the Formspring.me website make up the XML file known as the Formspring.me dataset. "Cyberbullying Positive" and "Cyberbullying Negative" are two categories within the dataset. Positive messages contain cyberbullying, whereas negative messages represent messages that do not. There are 12 266 messages in the class of cyberbullying negative and 892 messages in the class of cyberbullying positive. The messages collected from Myspace group chats make up the Myspace dataset. The group chats in the dataset have labels and are arranged into ten message groups. If there are 100 messages in a group conversation, the first group will consist of 1–10 messages, the second group will consist of 2–11 messages, and the final group will consist of 91–100 messages. Each group of ten messages receives a single label, indicating whether or not bullying is present in those ten messages. The 1753 message groups in this dataset are divided into 10 groups, each with 357 labels that are positive (bullying) and 1396 labels that are negative (non-bullying).

From Twitter, Atoum [140] gathered two datasets (Dataset-1 and Dataset-2) one month apart. Twitter dataset 1 contains 6463 tweets, 2521 of which are about cyberbullying and 3942 are not. 3721 tweets make up Twitter dataset 2, of which 1374 are about cyberbullying and 2347 are not.

TABLE III. CYBERBULLYING DATASETS

Dataset	Category	Number of Classes	Classes	Social Network Platform	Size	Availability	Year
Mangaonkar et al. [125]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	1340	N/A	2015
Waseem and Hovy [126]	Cyber Threats and Harassment	3	Racism Sexism None	Twitter	16 K	[141]	2016
Zhao et al [127]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	1762	N/A	2016
Singh et al. [128]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	4865	N/A	2016
Al-garadi et al [130]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	10,007	N/A	2016
Hosseinmardi et al. [131]	Flaming and Stalking and Harassment	2	Bullying Non-Bullying	Instagram	1954	N/A	2016
Zhang et al. [132]	Trolling and Harassment	2	Bullying Non-Bullying	Formspring	13 K	N/A	2016
Wulczyn et al. [133]	Denigration and Masquerade and Harassment	2	Attacking Non-Attacking	Wikipedia	100 K	[142]	2017

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
Article Received: *25 July 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *06 October 2023*

Davidson et al. [134]	Trolling and Harassment	3	Bullying Non-Bullying Neither	Twitter	33,458	[143]	2017
Founta et al. [135]	Cyber Threats and Harassment	7	Offensive Abusive Hateful Aggressive Cyberbullying Spam Normal	Twitter	100 K	[144]	2018
De Gibert et al. [136]	Trolling and Harassment	2	Hateful Non-Hateful	Stormfront	10,568	[145]	2018
Banerjee et al. [137]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	69,874	N/A	2019
Sadiq et al. [138]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter	20,001	[146]	2021
Kumar and Sachdeva [139]	Trolling and Harassment	2	Bullying Non-Bullying	Formspring MySpace	13,158 1753	N/A	2022
Atoum [140]	Trolling and Harassment	2	Bullying Non-Bullying	Twitter Dataset 1 Twitter Dataset 2	6463 3721	N/A	2023

The vast majority of studies and experiments used Twitter datasets, as shown in Table 3. This is because tweets can be easily accessed and made available utilizing the Twitter API to crawl them. Out of everything, the majority of research focuses on identifying cyberbullying and separating it from non-cyberbullying (offensive) texts.

C. Cyberbullying detection approaches

To detect and identify abusive language, a few sentiment-based methods have been published recently [147]. These methods are the machine-learning method, the lexicon-based method and the hybrid method, as shown in Figure 8.

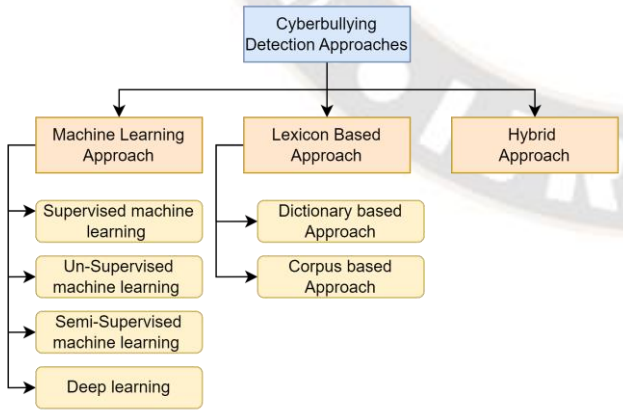

Figure 8. Cyberbullying Detection Approaches

The Machine Learning Approach (MLA) consists of the following methods: supervised machine learning, un-supervised machine learning, semi-supervised machine learning and deep learning. A classifier is created in a supervised machine learning approach to automatically learn the characteristics of categories or classes from a set of pre-annotated training textual content. Some major issues and challenges must be considered when using the supervised machine learning approach, including the categories utilized for classifying the instances, the labelled training data, the retrieved and selected features utilized to represent every unknown textual content, and the chosen algorithm utilized for categorization [148]. Unsupervised machine learning seeks to uncover and comprehend the hidden structure in unlabeled data [149]. In semi-supervised learning, the behaviour of learning is examined with respect to the combination of labelled and unlabeled data, and algorithms that benefit from this combination are developed [150]. A developing area of machine learning is called "Deep Learning," which draws inspiration from artificial neural networks [151]. It provides supervised and unsupervised methods for learning data representations with the help of the hierarchy of layers, enabling multiple processing [152].

Lexicon-Based Approach (LBA) involves the construction of a dictionary from which words are searched and counted in text. For the purpose of classifying textual content, these calculated frequencies can be used explicitly as features or as scores. The use of domain-specific words in a dictionary may limit the effectiveness of this approach with regards to classification; Also, the manual scoring of domain-specific words needs to be automated To minimize the amount of manpower needed [153]. The corpus-based approach recognizes new sentiment words and their polarity from a large collection of sentiment words with pre-defined polarities [154]. Incorporating sentiment labels and context into data-driven

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
Article Received: *25 July 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *06 October 2023*

approaches gives access to not only sentiment labels but also context. The dictionary-based approach makes use of the lexicographical tools such as Artha (https://sourceforge.net/projects/artha/, access on 16 February 2023), Tematres (https://www.vocabularyserver.com/, access on 16 February 2023), Wordhoard (https://wordhoard.northwestern.edu/, access on 16 February 2023), or WordNet (https://wordnet.princeton.edu/, access on 16 February 2023). A key strategy in this process involves gathering an initial collection of sentiment words, manually orienting them; then, using a dictionary to find synonyms and antonyms of these words to enlarge this collection [155].

As a final approach, the Hybrid Approach (HA) integrates lexicon-based and machine learning methods.

D. Cyberbullying detection techniques

Research on text mining and analysis has become increasingly popular and active. Such data are widely accessible, which makes text analytics an important factor. Based on the number of classes in these datasets, cyberbullying detection tasks can be accomplished as binary or multi-class classifications.

1. Binary cyberbullying classification

Cyberbullying detection has been examined as a binary classification task, such as "Hate-vs.-Non-Hate" or "cyberbullying-vs.-non-cyberbullying." An overview of studies using binary classification methods for cyberbullying is provided in this section.

Mangaonkar et al. [125], in order to categorise tweets, used various types algorithms, followed by AND and OR parallelism. In order to enhance performance, they combined the output from various classifiers. They conducted experiments with homogeneous (all computing nodes use a single classification algorithm), heterogeneous (each node uses a distinct algorithm), and selective (the best-performing node is selected as the expert, and all other nodes defer to it) collaborations to categorise tweets using a four-node detection system. Each tweet is examined by every node and labelled as cyberbullying if more than 50% of the nodes in the AND configuration or any node in the OR configuration flag it as such. They found that whereas AND parallelism yields the highest accuracy at 70%, OR parallelism yields the highest recall values at 60%.

Nandhini and Sheeba [156] proposed a system for identifying online bullying activity on social networks in the English language, to aid the government in taking action before more people become victims of cyberbullying. Dataset used consists of nearly 4K records, which was collected from social networks (Formspring.me and Myspace.com) [90]. They employed the Naive Bayes (NB) classifier for this, which achieved 92% accuracy on the Form-spring dataset and 91% accuracy on the MySpace.me dataset.

Zhao et al. [127], suggested Embedding-enhanced Bag-of-Words (EBoW), a novel representation learning technique for cyberbullying detection. EBoW combines bullying characteristics, latent semantic features, and bag of words features. Word embeddings, which are capable of capturing the semantic information contained in words, are utilized to generate the characteristics of bullying. A linear SVM with a recall of 79.4 is applied to detect bullying messages after the final representation is learned.

Singh et al. [128] combined social and text information as the classifier's input, using probabilistic fusion approaches. The English Twitter dataset has been used to test the suggested methodology. The obtained results' accuracy was 89%.

Al-garadi et al. [130], to identify cyberbullying on Twitter in the English language, used supervised machine learning algorithms like NB, SVM, K Nearest Neighbour (KNN) and Random Forest (RF). The model's accuracy is 70.4% by NB, 50% by SVM, 56.8% by KNN and 62.9% by Random Forest (RF), according to an evaluation.

Hosseinmardi et al. [131], explored the issue of anticipating cyberbullying in the social network based on the Instagram media. With a Logistic Regression (LR) classifier achieving 72% recall and 78% precision, they showed that non-text features like image and user metadata were significant in predicting cyberbullying.

Zhang et al. [132] proposed a brand-new Pronunciation-based Convolutional Neural Net- work (PCNN), for the purpose of identifying cyberbullying. They used an English-language cyberbullying dataset from Formspring.me to evaluate the efficacy of their model. In their study, they found that PCNN is capable of achieving 88.1% accuracy.

Wulczyn et al. [133] proposed a methodology in cyberbullying detection, by using LR and Multi-Layer Perceptron (MLP) on Wikipedia, producing an open dataset of over 100 k high-quality human-labeled comments. Area Under the Receiver Operating Characteristic (AUROC) was used to assess their models, and they achieved 96.18% using LR and 96.59% using MLP.

De Gibert et al. [136], conducted thorough qualitative and quantitative analysis of their dataset along with the classification models SVM, Convolution Neural Networks (CNN), and Long-Short Term Memory (LSTM). A well-balanced subset of labelled sentences is used in the experiments. A total of 2 k labelled sentences—all the HATE sentences plus an equal number of NOHATE sentences—were collected. Eighty percent of this sum was allocated for training, and the remaining twenty percent was allocated for testing.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

SVM, CNN, and LSTM, the evaluated algorithms, respectively achieved 71%, 66%, and 73% accuracy.

Banerjee et al. [137], proposed an approach for the detection of cyberbullying in the English language. They used CNN on a set of 69,874 tweets from Twitter. Their suggested method had a 93.97% accuracy rate.

Sadiq et al.'s system [138] used English tweets from the cyber-troll dataset, for detecting cyberbullying by employing CNN with LSTM and CNN with BiLSTM. Statistical findings demonstrated that the model they suggested detects aggressive behaviour with 92% accuracy.

Kumar and Sachdeva [139] proposed the Bi-GRU Attention-CapsNet (Bi- GAC) hybrid model, for the purpose of detecting cyberbullying in social media text, which gains from learning sequential semantic representations and spatial location data utilising a Bi-GRU with self-attention followed by CapsNet. The F1-score and the ROC-AUC curve are used as performance metrics to assess the proposed Bi-GAC model. The results perform better than currently used methods on the benchmark Formspring.me and MySpace datasets. The F1-score for the datasets from MySpace and Formspring.me outperformed traditional models by nearly 94% and 93%, respectively.

Atoum [140] developed an effective sentiment analysis and language modeling technique to identify cyberbullying in tweets. On the basis of two tweet datasets, various machine learning algorithms are compared and contrasted. Other ML classifiers like DT, RF, NB, and SVM were outperformed by CNN classifiers using larger n-gram language models. CNN classifiers had an average accuracy of 93.62% and 91.03%.

2. Multi-class cyberbullying classification

Numerous studies have been carried out to classify cyberbullying into multiple categories. In this section, the studies on multi-class cyberbullying classification methods are compiled.

Waseem and Hovy [126] examined the efficacy of various features in the classification of cyberbullying. With an F1-score of 73%, they tested and measured the effect of different features on prediction performance using an LR classifier and 10-fold cross-validation.

Badjatiya et al. [157] used deep neural network architectures for hate speech detection in the English language [126]. They suggested combining gradient-boosted decision trees with deep neural network model embeddings to achieve higher accuracy levels. Combining deep neural network models with embeddings and gradient-boosted decision trees produced the best accuracy results with a F1-score of 93%.

Park and Fung [158] for the purpose of detecting and identifying sexist and racist languages, proposed a two-step method of abusive language classification. They first categorise the language as abusive or not, and in a subsequent step, they categorise it into explicit types. Their method displays an impressive result of 82.7% F1-score using Hybrid-CNN in the first step and 82.4% F1-score using LR in the second step with a public English Twitter corpus [126] comprising 20 thousand tweets of a sexist and racist nature.

Watanabe et al. [159] proposed a method to identify hate speech on Twitter in the English language [126]. The suggested method subsequently identifies hate speech patterns and signs by extracting features from unigrams along with sentimental and semantic features to categorize tweets as hateful, offensive, or clean. The proposed method successfully classifies tweets with an accuracy of 78.4%.

Wang et al. [160] suggested a framework for Metamorphic Testing for Textual Content Moderation (MTTM) software. 2000 text messages from actual users were used in a pilot study, and eleven metamorphic relations were summarized at the character, word, and sentence levels of perturbation. To create test cases that are still harmful but are unlikely to be moderated, MTTM applies these metamorphic relations to the harmful textual content. When the MTTM is put to the test, the results reveal that it can achieve error-finding rates of up to 83.9%.

3. Comparative analysis of the literature review

Table 4 summarizes the dataset used in the experimentation and the number of classes, the algorithms tried, and the outcomes for each of the binary and multiclass classification works previously described.

According to Table 4 of this comparative study, binary classification is the most frequently performed task in cyberbullying detection.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
Article Received: *25 July 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *06 October 2023*

TABLE IV. CARTOGRAPHY OF EXISTING RESEARCH IN CYBERBULLYING DETECTION

Author	Classes	Dataset	Approach	Algorithm	Evaluation Metric	
Mangaonkar et al. 2015 [125]	2 Classes (Cyberbullying, Non-Cyberbullying)	Twitter	MLA	LR (OR parallelism)	Recall	60%
				LR (AND parallelism)	Accuracy	70%
Nandhini and Sheeba, 2015 [156]	2 Classes (Cyberbullying–NonCyberbullying)	Formspring	MLA	NB	Accuracy	92%
		MySpace.com				91%
Waseem and Hovy 2016 [126]	3 Classes (Sexism, Racism, Neither)	Twitter	MLA	LR	F1-score	73%
Zhao et al. 2016 [127]	2 Classes (Cyberbullying, Non-Cyberbullying)	Twitter	MLA	SVM	F1-score	79.4%
Singh et al. 2016 [128]	2 Classes (Cyberbullying, Non-Cyberbullying)	Twitter	LBA	Probabilistic Fusion approach	Accuracy	89%
Al-garadi et al. 2016 [130]	2 Classes (Cyberbullying, Non-Cyberbullying)	Twitter	MLA	NB	Accuracy	70.4%
				SVM		50%
				RF		62.9%
				KNN		56.8%
Hosseinmardi et al. 2016 [131]	2 Classes (Cyberbullying, Non-Cyberbullying)	Instagram	MLA	LR	Recall	72%
					Precision	78%
Zhang et al. 2016 [132]	2 Classes (Cyberbullying, Non-Cyberbullying)	Formspring	MLA	PCCN	Accuracy	88.1%
Wulczyn et al. 2017 [133]	2 Classes (Attacking, Non-Attacking)	Wikipedia	MLA	LR	AUROC	96.18%
				MLP		96.59%
Badjatiya, Pinkesh et al. 2017 [151]	3 classes (Sexism, Racism, Neither)	Twitter	MLA	LSTM	F1-score	93%
Park, Ji Ho et al. 2017 [158]	3 classes (Sexism, Racism, Neither)	Twitter	MLA	CNN	F1-score	82.7%
				LR		82.4%
De Gibert et al. 2018 [136]	2 Classes (Hate, Non-Hate)	Stormfront	MLA	SVM	Accuracy	71%
				CNN		66%
				LSTM		73%
Watanabe et al. 2018 [159]	3 Classes (Hateful, Offensive and Clean)	Twitter	MLA	J48graft	Precision	88%
					Recall	87.4%
					F1-score	87.5%
Banerjee et al. 2019 [137]	2 Classes (Cyberbullying–NonCyberbullying)	Twitter	MLA	CNN	Accuracy	93.97%
Sadiq et al. 2021 [138]	2 Classes (Cyber-aggressive, Non-Cyber-aggressive)	Twitter	MLA	CNN + LSTM + Bi-LSTM	Accuracy	92%
Kumar and Sachdeva 2022 [139]	2 Classes (Cyberbullying–Non-Cyberbullying)	Formspring	HA	Bi-GAC	F1-score	94.03%
		MySpace				93.89%
Wang et al. 2023 [161]	3 Classes (Cyberbullying–Non-Cyberbullying, Neither)	Twitter	HA	MTTM	Error Finding Rates	83.9%
Atoum, 2023 [140]	2 Classes (Cyberbullying–Non-Cyberbullying)	Twitter Dataset 1	MLA	CNN	Accuracy	93.62%
		Twitter Dataset 2				91.03%

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
Article Received: *25 July 2023* ***Revised:*** *18 September 2023* ***Accepted:*** *06 October 2023*

As was already mentioned, binary classification, as opposed to multi-class classification, is the most frequently performed task in cyberbullying detection. Cyberbullying texts are regarded as examples of the "bullying" category, while all other texts fall under the "non-bullying" category. Compared with other social media platforms, Twitter is the data source that is most frequently studied. The majority of researchers used and compared numerous supervised machine learning algorithms to identify the most effective ones for problems with cyberbullying detection. SVM has been used to create cyberbullying prediction models and has been discovered to be accurate and effective when compared to the conventional machine learning algorithms. Deep learning algorithm that is most popular for binary or multiple-class classification of cyberbullying is CNN. Researchers use a variety of evaluation metrics, such as the F1-score, accuracy, recall, and Precision, to assess how well their proposed model performs in separating cyberbullying texts from non-cyberbullying texts [162,163]. After that, the binary and multi-class classification algorithms are examined in light of the outcomes they produced.

Figure 9 shows the accuracy of binary cyberbullying classification on different English datasets. It shows that compared to SVM, NB, CNN + LSTM, and CNN + LSTM + Bi-LSTM, CNN has higher accuracy. The accuracy provided by NB on various datasets ranges from 91% to 92%, which is also acceptable.

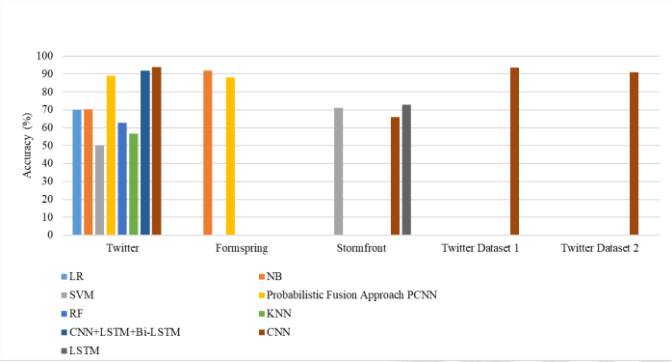

Figure 9. Accuracy of Binary Cyberbullying Classification in English

Figure 10 shows the F1-Score of multiple class cyberbullying classifications on Twitter in English. It demonstrates how LSTM, which had an F1-Score of 93%, outperformed all other machine learning algorithms used.

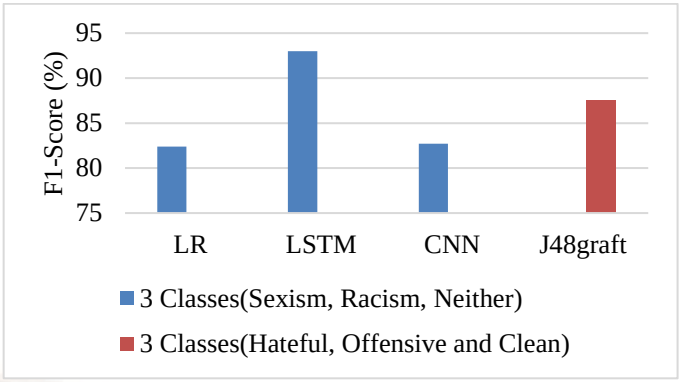

Figure 10. F1-Score of Multiple Class Cyberbullying Classification on Twitter in English

4. Taxonomy of cyberbullying detection techniques

Numerous researchers are working to investigate every possible angle of cyberbullying detection. We have created the following taxonomy after taking into account all the methods examined during the literature review. The taxonomy depicted in figure 11 provides a clear picture of the various proposed techniques for detecting cyberbullying.

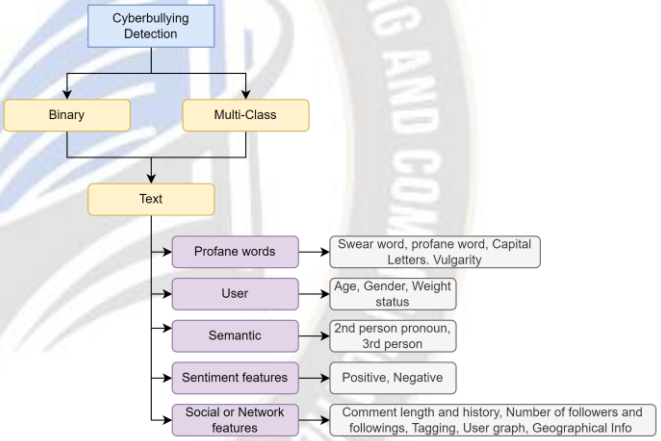

Figure 11. Taxonomy of cyberbullying detection techniques

V. FACTORS, PREVENTIVE MEASURES AND LAWS TO COMBAT CYBERBULLYING BEHAVIOR

This section briefly outlines the motivating factors that lead people to engage in cyberbullying behaviour, potential preventive measures to avoid it, and the general legal frameworks of various nations to address them. The following is a discussion of some of the reasons of cyberbullying behaviour in people.

(1) *Psychological Disorders:* Most often, those who engage in cyberbullying behaviour may already be suffering from severe mental health problems brought on by the bullying they experienced.

(2) *Personality Qualities:* Few people may possess a dark tetrad personality, which tends to lack empathy for others and enjoy bullying others because it gives them

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

a sense of power.

(3) *Repercussions of persistent bullying:* People who have experienced bullying have a propensity to engage in offensive behaviour as a form of retaliation.

(4) *Repercussions of Disputes or Breakups:* Conflicts that lead to the end of a friendship or relationship can occasionally breed resentment and jealousy, which encourages harassing behaviour online.

(5) *Boredom or a desire to experiment with new things:* "An idle mind is a devil's workshop," as the proverb goes, cyberbullying is a behaviour used by people who are bored or have nothing to do to experiment with a new persona.

(6) *Isolation or Loneliness:* When someone feels alone or ignored by others, they may act in an offensive manner as a way to lash out or vent their rage.

(7) *Anonymity and Non-Confrontation the Internet Offers:* When done anonymously, cyberbullying becomes non-confrontational behaviour. Additionally, this anonymity may encourage some people to post offensive content online.

(8) *Absence of Need for Popularity or Physical Dominance:* No matter how someone is in real life, they can engage in offensive behaviour online.

(9) *Simple Accessibility:* Anyone with internet access is capable of acting inappropriately towards others. The majority of the time, known individuals only bully, making it simple for them to contact the target online.

(10) *No Response from the Victim:* Offensive behaviour that occurs offline enables the perpetrator to see the results of their actions and may persuade them to back off. However, online offenders won't be able to see it and carry on with their behaviour for a while [164].

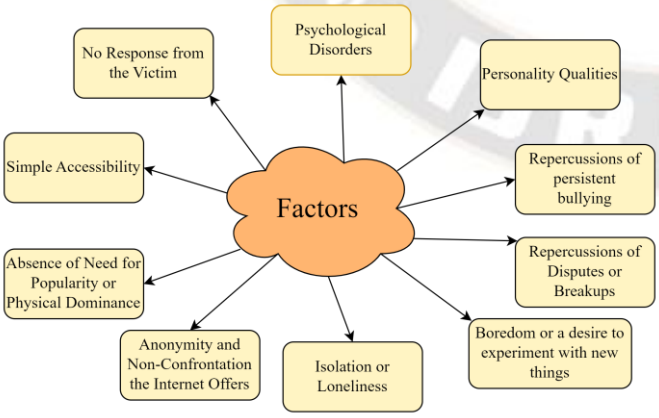

Figure 12. Factors that drive Individuals to Engage in Cyberbullying Behaviour

Figure 12 depicts the factors that drive Individuals to Engage in Cyberbullying Behaviour. A crucial aspect of the task of preventing cyberbullying is the potential to detect cyberbullying behaviour. The only goal of identifying cyberbullying behaviour is to stop it from occurring. Most of the research literature we looked at in our survey only addressed their detection, which can then be used to prevent them from occurring. Bullies can be prevented from engaging in toxic online behaviour by addressing a few issues that led them to engage in these offensive behaviours. Instead of dealing with these problems alone, the first step in this regard is to discuss them openly with anyone who has already dealt with them. People who have experienced cyberbullying frequently act offensively towards others as a way to vent their anger. To prevent this, a space where people can openly discuss their experiences and let go of their internal burdens can be created. By educating people thoroughly through all forms of media about the cyberlaws of the specific country and how serious of a crime it is, as well as the severity of the punishment one might have to face if they get caught in the act, it is possible to address factors like feelings of isolation and boredom that make people indulge in exhibiting cyberbullying behaviour.

In general, one should ensure that all of their electronic devices are password-protected and shouldn't be left unattended, especially when in a public or unfamiliar setting. Below are a few additional preventative measures that can be used.

(1) *Utilise the privacy setting to the fullest extent possible:* The majority of social media platforms allow users to change their privacy settings. Users can control who has access to their posts and personal data. Therefore, by using it, one can shield themselves from cyberbullies [165].

(2) *Post content only after carefully considering it:* Public platforms include social media sites. What is posted there cannot be deleted because even though it can be taken down from the platform, what if a copy was saved before it was deleted? The content is then independent of the person who posted it. So, it's important to consider your options before posting sensitive data such as photos or videos *etc*., [166].

(3) *Avoid retaliating:* The majority of the time, cyberbullies push themselves to attract the target's attention. They constantly attempt to make things worse. One will be giving them what they just wanted if they take revenge. Regardless of how offensive or untrue their post is, you can decide to ignore it. Do not try to defend it or respond to them in any way [167].

(4) *Educate oneself and others about the effects and behaviours of cyberbullying:* Understanding the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

behaviours that constitute cyberbullying can help prevent acting out and escalating any incident. Learning about the effects and spreading that information to others could stop many people from unintentionally engaging in cyberbullying [167].

Popular social media sites like Twitter give users the option to report offensive content if they discover it there. They do take a quick look at the content that has been reported, and if they, too, find it offensive, they remove them right away. Additionally, it temporarily suspends that user's account. If they publish more harmful material, their account will be permanently suspended. However, this procedure cannot start right away after the offensive material is published. Therefore, there is an urgent need to create a model that accurately detects objectionable content and stops it from being posted on social media in real time. The judicial procedures used in various nations to address the issue of cyberbullying behaviour vary. Table 5 shows the penalties for various forms of cyberbullying behaviour in a few nations with strict laws to prevent it [168].

TABLE V. LAWS OF DIFFERENT NATIONS TO COMBAT CYBER OFFENSIVE BEHAVIOR

Country	Type of offensive Behaviour	Under the Act
Canada	Cyberbullying	Education Act
United Kingdom	Cyberbullying	Malicious Communications Act
USA (Hawaii state)	Cyberbullying	SB2094 Law
USA (Louisiana state)	Cyberbullying	H.B.1259 Act, 989
USA(Maryland)	Cyberbullying	Grace's law
USA (North Carolina)	Cyberbullying	14-458.1
USA(Tennessee)	Cyberbullying, Online threatening	SB
India	Cheating using computer resources	IT Act 2000 Section 66(D)
India	Violating privacy digitally (sharing pictures or information online without consent)	IT Act 2000 Section 66(E)
India	Uploading, circulating, transferring offensive, vulgar materials online	IT ACT 2000 Section 67
India	Intimidating someone anonymously	IPC 1860 Section 507
India	Online stalking with an intension to hurt	IPC 1860 Section 354(D)

VI. CYBERBULLYING CHALLENGES AND LIMITATIONS

In this work, several problems that have an impact on the majority of the recent research on detecting cyberbullying were identified:

- Data scarcity.
- Contextual uncertainty.
- Accessibility and availability of social network data.
- Manual Data Labelling.
- The level of cyberbullying severity.

Owing to the difficulty of gathering reliable data on cyberbullying in the wild, the field experiences a lack of data in various languages. Another difficulty is figuring out the context of a conversation. It's important to consider the context because many words are, in fact, ambiguous. Current methods for detecting cyberbullying rely on potential victims' experiences and the availability of precise, pertinent information from social media accounts. However, in practice,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

social network restrictions and consumer privacy practices have an impact on the data's accessibility. Maintaining privacy is regarded as a difficult issue. A suitable solution necessitates that the person understands their privacy preferences. Data labelling is a labor-intensive and time-consuming process because, before the process even begins, it is necessary to choose the appropriate definitions of key terms that will be used during the labelling of ground truth. The severity of cyberbullying is thought to be difficult to assess. In addition to machine learning understanding, predicting the different levels of cyberbullying severity requires a thorough analysis to identify and classify levels of cyberbullying severity from social and psychological experiences.

VII. FUTURE RESEARCH DIRECTIONS

This section outlines the potential future research areas that the research community should focus on in order to combat the issue of cyberbullying.

(1) *Automate annotating method*

As part of the classification of cyberbullying, annotation is a crucial step in labeling the data. Data annotations in most cyberbullying identification tasks are manual, so there is a high chance that they will be subjective. In order to avoid bias and make annotations more objective, an automated annotation model should be developed.

(2) *Significance of the writing style of an individual*

There is a unique style of writing or typing for each individual. The cyberbullying detection task may be enhanced if these techniques are identified and utilized along with a variety of embedding techniques.

(3) *Determining multimedia offensive content efficiently*

Social media today widely uses videos, audio, gifs, and pictures, and it is essential to remove those that contain offensive or abusive content in addition to text. According to our survey, most of the works are concerned with textual cyberbullying content identification and very few with other multimedia toxic content. The performance of these works is not satisfactory and should be improved significantly even though few researchers are working to identify abusive multimedia content.

(4) *Determining the source of cyberbullying content*

Cyberbullying can only be prevented by understanding why and how it started, beyond identifying the victim and bully. There has been little attention paid to this so putting some effort into this area and combining it with other techniques might make a difference.

(5) *Identify and remove the cyberbullying content in real-time*

According to our survey, most of the work on detecting cyberbullying content on social media has been done on the dataset collected so far. In order to understand social media posts in real-time, a model must be developed. Additionally, it should be able to block any content that is explicitly or implicitly cyberbullying as soon as it is discovered.

(6) *Developing a generalized all in one model for varied social media platforms*

The majority of research studies only concentrate on the data gathered from one specific social media platform, such as Twitter (tweets), Instagram (images), etc. To easily solve the issue of identifying cyberbullying content that is posted on social media, it is crucial to develop a generic model to identify all types of cyberbullying content, including text, images, gifs, video, etc., in various social media platforms in real-time.

(7) *Identification of the seriousness of the cyberbullying*

Even though there has been numerous works on identifying textual bullying content, most of it has not been successful in determining the severity of the bullying behaviour. Once the severity is established, appropriate countermeasures can be implemented.

(8) *Develop images, audios, videos containing cyberbullying content datasets*

Our survey indicates that the lack of datasets is the primary reason of the limited work on non-textual content. As sharing audio, video, and images on various social network platforms is on the rise, especially among adolescents, the research community should concentrate on creating and quickly processing those datasets.

(9) *Developing detection algorithm for regional languages*

In most social media platforms, users can post content in a variety of regional languages due to the technological advancements. The majority of offensive language detections have been in English and a few other foreign languages. Hence, it is crucial to create datasets, especially for the specific regional languages, and to create algorithms to effectively identify the offensive content posted in that language.

(10) *Incorporate psychological aspects along with the other features into the detection models*

The most of the studies looking into the issue of finding cyberbullying content downplay the significance of the bully's psychology or mental state. Work together with psychologists and sociologists to improve the accuracy and effectiveness of the detection models. Try incorporating their recommendations into the detection model in addition to other features that are already in

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

place, as this may help to more fully understand the psychology of bullies and make it easier to identify them.

(11) *Assimilate temporal aspects along with the network-based features*

Our survey indicates that the majority of the work has overlooked the significance of temporal factors in identifying cyberbullying content. They could be combined with network- or graph-based features to provide more information and boost the effectiveness of the detection model.

VIII. CONCLUSIONS

Online offensive/cyberbullying behaviour is a serious crime. A victim may end up committing suicide. It is therefore crucial to detect them and eliminate them as soon as possible. A review of existing literature on detecting cyberbullying behavior on social media platforms using a variety of techniques is provided in this paper.

English-language existing cyberbullying datasets have been examined. We conducted a thorough analysis of the evaluation procedures, dataset size, and dataset source used in the most recent studies in this area. A comparative study that incorporates binary and multiple class cyberbullying classification has also been introduced, summarizing the most recent work that has been done over the past few years. There is discussion of the causes of cyberbullying, preventative measures to be taken, and global laws to stop this kind of online misbehaviour. Finally, a thorough description of the main challenges and open research issues was provided.

REFERENCES

[1] Omuya, E. O., Okeyo, G., & Kimwele, M. (2023). Sentiment analysis on social media tweets using dimensionality reduction and natural language processing. Engineering Reports, 5(3), e12579.

[2] https://datareportal.com/reports/digital-2023-global-overview-report

[3] Available online: https://www.Oberlo.Com/Statistics/How-Many-People-Use-Social-Media (accessed on 16 April 2023).

[4] https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[5] Corazza, M.; Menini, S.; Cabrio, E.; Tonelli, S.; Villata, S. A Multilingual Evaluation for Online Hate Speech Detection. ACM Trans. Internet Technol. 2020, 20, 1–22.

[6] StopBullying.Gov. Available online: https://www.stopbullying.gov.

[7] Bisht, A.; Singh, A.; Bhadauria, H.S.; Virmani, J. Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. Recent Trends Image Signal Process. Comput. Vis. 2020, 1124, 243–264.

[8] Blaya, C. Cyberhate: A Review and Content Analysis of Intervention Strategies. Aggress. Violent Behav. 2018, 45, 163–172.

[9] Chinivar, S., Roopa, M. S., Arunalatha, J. S., & Venugopal, K. R. (2022). Online offensive behaviour in socialmedia: Detection approaches, comprehensive review and future directions. Entertainment Computing, 100544.

[10] Wikipedia, Hate Speech Definition. https://en.wikipedia.org/wiki/Online_hate_speech.

[11] Wikipedia, Misogyny Definition. https://en.wikipedia.org/wiki/Misogyny.

[12] COLLINSDICTIONARY, Xenophobia Definition. https://www.collinsdictionary.com/dictionary/english/xenophobia.

[13] I. Wigmore, Troll Definition. https://whatis.techtarget.com/definition/trolling/.

[14] I.G. DICTIONARY, Cyber Aggression Definition. https://www.igiglobal.com/dictionary/cyber-aggression/6573/.

[15] P.H. TEAM, Cyber bullying definition, 2017, https://blog.providence.org/archive/cyber-aggression-vs-cyberbullying-and-how-to-keep-your-child-safe.

[16] B. Joseph, Cyber bullying definition, 2018, https://kidshealth.org/en/teens/cyberbullying.html.

[17] Stopbullying.gov, Cyber Bullying Definition. https://www.stopbullying.gov/cyberbullying/what-is-it/.

[18] A. INGHAM, Cyber bullying instance, 2018, https://www.familyorbit.com/blog/real-life-cyberbullying-horror-stories/.

[19] J. Raskauskas, A.D. Stoltz, Involvement in traditional and electronic bullying among adolescents, Dev. Psychol. 43 (3) (2007) 564–575.

[20] K.R. Williams, N.G. Guerra, Prevalence and predictors of internet bullying, J. Adolesc. Health 41 (6) (2007) 14–21.

[21] M.L. Ybarra, K.J. Mitchell, Youth engaging in online harassment: associations with caregiver-child relationships, internet use, and personal characteristics, J. Adolesc. 27 (3) (2004) 319–336.

[22] S.D. Freis, R.A.R. Gurung, A facebook analysis of helping behavior in online bullying, Psychol. Pop. Media Cult. 2 (1) (2013) 11–19.

[23] R.M. Kowalski, S.P. Limber, P.W. Agatston, Cyberbullying: Bullying in the Digital Age, Wiley-Blackwell, 2012.

[24] J.A. Casas, R. Del Rey, R. Ortega-Ruiz, Bullying and cyberbullying: convergent and divergent predictor variables, Comput. Human Behav. 29 (3) (2013) 580–587.

[25] J.W. Patchin, S. Hinduja, Bullies move beyond the schoolyard, Youth Violence Juv. Justice 4 (2) (2006) 148–169.

[26] R.S. Tokunaga, Following you home from school: a critical review and synthesis of research on cyberbullying victimization, Comput. Human Behav. 26 (3) (2010) 277–287.

[27] Chan, T. K., Cheung, C. M., & Lee, Z. W. (2021). Cyberbullying on social networking sites: A literature

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

review and future research directions. Information & Management, 58(2), 103411.

[28] L.A. McFarland, R.E. Ployhart, Social media: a contextual framework to guide research and practice, J. Appl. Psychol. 100 (6) (2015) 1653–1677.

[29] S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. DeSmet, I. De Bourdeaudhuij, Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully, Comput. Human Behav. 31 (February 2014) (2014) 259–271.

[30] Azumah, S. W., Elsayed, N., ElSayed, Z., & Ozer, M. (2023). Cyberbullying in Text Content Detection: An Analytical Review. arXiv preprint arXiv:2303.10502.

[31] Hang, O.C.; Dahlan, H.M. Cyberbullying Lexicon for Social Media. In Proceedings of the Research and Innovation in Information Systems (ICRIIS), Johor Bahru, Malaysia, 2–3 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.

[32] Sangwan, S.R.; Bhatia, M.P.S. Denigration Bullying Resolution UsingWolf Search Optimized Online Reputation Rumour Detection. Procedia Comput. Sci. 2020, 173, 305–314.

[33] Colton, D.; Hofmann, M. Sampling Techniques to Overcome Class Imbalance in a Cyberbullying Context. Comput. Linguist. Res. 2019, 3, 21–40.

[34] Qodir, A.; Diponegoro, A.M.; Safaria, T. Cyberbullying, Happiness, and Style of Humor among Perpetrators: Is There a Relationship? Humanit. Soc. Sci. Rev. 2019, 7, 200–206. [CrossRef]

[35] Peled, Y. Cyberbullying and Its Influence on Academic, Social, and Emotional Development of Undergraduate Students. Heliyon 2019, 5, e01393.

[36] Dhillon, G.; Smith, K.J. Defining Objectives for Preventing Cyberstalking. Bus. Ethics 2019, 157, 137–158.

[37] la Vega, D.; Mojica, L.G.; Ng, V. Modeling Trolling in Social Media Conversations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, 7–12 May 2018; pp. 3701–3706.

[38] Hassan, S.; Yacob, M.I.; Nguyen, T.; Zambri, S. Social Media Influencer and Cyberbullying: A Lesson Learned from Preliminary Findings. In Proceedings of the 9th Knowledge Management International Conference (KMICe), Miri, Sarawak, Malaysia, 25–27 July 2018; pp. 200–205.

[39] Raisi, E.; Huang, B. Weakly Supervised Cyberbullying Detection Using Co-Trained Ensembles of Embedding Models. In Proceedings of the Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 479–486.

[40] Willard, N.E. Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress; Research Press: Champaign, IL, USA, 2007.

[41] L.A. McFarland, R.E. Ployhart, Social media: a contextual framework to guide research and practice, J. Appl. Psychol. 100 (6) (2015) 1653–1677.

[42] G.C. Kane, M. Alavi, G. Labianca, S.P. Borgatti, What's different about social media networks? A framework and research agenda, Mis Q. 38 (1) (2014) 274–304.

[43] J. Barlin´ska, A. Szuster, M. Winiewski, Cyberbullying among adolescent bystanders: role of the communication medium, form of violence, and empathy, J. Commun. Appl. Soc. Psychol. 23 (1) (2013) 37–51.

[44] T.K.H. Chan, C.M.K. Cheung, R.Y.M. Wong, Cyberbullying on social networking sites: the crime opportunity and affordance perspectives, J. Manag. Inf. Syst. 36 (2) (2019) 574–609.

[45] R.M. Kowalski, S.P. Limber, P.W. Agatston, Cyberbullying: Bullying in the Digital Age, Wiley-Blackwell, 2012.

[46] D. Olweus, Bullying at School: What We Know and What We Can Do (understanding Children's Worlds), Blackwell, Malden, MA, 1993.

[47] H.J. Thomas, J.P. Connor, J.G. Scott, Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents – a review, Educ. Psychol. Rev. 27 (March 2015) (2015) 135–152.

[48] P.K. Smith, G. Steffgen, Cyberbullying Through the New Media: Findings From an International Network, Psychology Press, 2013.

[49] E. Menesini, A. Nocentini, Definitions of cyberbullying. Cyberbullying Through the New Media, Psychology Press, 2013, pp. 41–54.

[50] E. Menesini, A. Nocentini, M. Camodeca, Morality, values, traditional bullying, and cyberbullying in adolescence, Br. J. Dev. Psychol. 31 (1) (2013) 1–14.

[51] R. Slonje, P.K. Smith, Cyberbullying: another main type of bullying? Scand. J. Psychol. 49 (2) (2008) 147–154.

[52] T. Vaillancourt, P. McDougall, S. Hymel, A. Krygsman, J. Miller, K. Stiver, C. Davis, Bullying: Are researchers and children/youth talking about the same thing? Int. J. Behav. Dev. 32 (6) (2008) 486–495.

[53] J. Pyz˙alski, From cyberbullying to electronic aggression: typology of the phenomenon, Emot. Behav. Difficulties 17 (3–4) (2012) 305–317.

[54] M. Rachoene, T. Oyedemi, From self-expression to social aggression: cyberbullying culture among south african youth on facebook, Communication 41 (3) (2015) 302–319.

[55] R. Dredge, J. Gleeson, X. de la Piedad Garcia, Cyberbullying in social networking sites: an adolescent victim's perspective, Comput. Human Behav. 36 (July 2014) (2014) 13–20

[56] H. Cowie, C.-A. Myers, Bullying amongst university students in the uk, Int. J. Emot. Educ. 6 (1) (2014) 66–75.

[57] T.R. Nansel, M. Overpeck, R.S. Pilla, W.J. Ruan, B. Simons-Morton, P. Scheidt, Bullying behaviors among us youth: prevalence and association with psychosocial adjustment, JAMA 285 (16) (2001) 2094–2100.

[58] W. Cassidy, C. Faucher, M. Jackson, Cyberbullying among youth: a comprehensive review of current international research and its implications and application to policy and practice, Sch. Psychol. Int. 34 (6) (2013) 575–612.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

[59] S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. DeSmet, I. de Bourdeaudhuij, 'Can i afford to help?' how affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites, Behav. Inform. Technol. 34 (4) (2015) 425–435.

[60] A. Bandura, Social cognitive theory of mass communication. Media Effects, Routledge, 2009, pp. 110–140.

[61] P.B. Lowry, J. Zhang, C. Wang, M. Siponen, Why do adults engage in cyberbullying on social media? An integration of online disinhibition and deindividuation effects with the social structure and social learning model, Inf. Syst. Res. 27 (4) (2016) 962–986.

[62] S. Camacho, K. Hassanein, M. Head, Cyberbullying impacts on victims' satisfaction with information and communication technologies: the role of perceived cyberbullying severity, Inf. Manag. 55 (4) (2018) 494–507.

[63] N. Brody, A.L. Vangelisti, Bystander intervention in cyberbullying, Commun. Monogr. 83 (1) (2015) 1–26.

[64] J. Anderson, M. Bresnahan, C. Musatics, Combating weight-based cyberbullying on facebook with the dissenter effect, Cyberpsychol. Behav. Soc. Netw. 17 (5) (2014) 281–286.

[65] M. Weber, M. Ziegele, A. Schnauber, Blaming the victim: the effects of extraversion and information disclosure on guilt attributions in cyberbullying, Cyberpsychol. Behav. Soc. Netw. 16 (4) (2013) 254–259.

[66] R. Thornberg, L. Wa¨nstro¨m, J.S. Hong, D.L. Espelage, Classroom relationship qualities and social-cognitive correlates of defending and passive bystanding in school bullying in sweden: a multilevel analysis, J. Sch. Psychol. 63 (August 2017) (2017) 49–62.

[67] S. Wachs, A. Go¨rzig, M.F. Wright, W. Schubarth, L. Bilz, Associations among adolescents' relationships with parents, peers, and teachers, self-efficacy, and willingness to intervene in bullying: a social cognitive approach, Int. J. Environ. Res. Public Health 17 (2) (2020) 420–436.

[68] W. Troop-Gordon, C.A. Frosch, C.M.W. Totura, A.N. Bailey, J.D. Jackson, R.D. Dvorak, Predicting the development of pro-bullying bystander behavior: a short-term longitudinal analysis, J. Sch. Psychol. 77 (December 2019) (2019) 77–89.

[69] D.J. Meter, S. Bauman, Moral disengagement about cyberbullying and parental monitoring: effects on traditional bullying and victimization via cyberbullying involvement, J. Early Adolesc. 38 (3) (2018) 303–326.

[70] L. Chen, S.S. Ho, M.O. Lwin, A meta-analysis of factors predicting cyberbullying perpetration and victimization: from the social cognitive and media effects approach, New Media Soc. 19 (8) (2016) 194–1213.

[71] B.S. Xiao, Y.M. Wong, Cyber-bullying among university students: an empirical investigation from the social cognitive perspective, Int. J. Bus. Inform. 8 (1) (2013) 34–69.

[72] K. Bussey, S. Fitzpatrick, A. Raman, The role of moral disengagement and self- efficacy in cyberbullying, J. Sch. Violence 14 (1) (2015) 30–46.

[73] D. Mann, Emotional Troubles for 'Cyberbullies' and Victims. WebMD Health News, 6 July 2010. http://www.webmd.com/parenting/news/20100706/emotional-troublesfor-cyberbullies- and-victims. Accessed 24 Aug 2015

[74] T. Wiguna, I.R. Ismail, R. Sekartini, N.S.W. Rahardjo, F. Kaligis, A.L. Prabowo, R. Hendarmo, The gender discrepancy in high-risk behaviour outcomes in adolescents who have experienced cyberbullying in Indonesia. Asian J. Psychiatry 37 (2018) (Elsevier)

[75] C. Nixon, Current perspectives: the impact of cyberbullying on adolescent health, in Adolescent Health, Medicine and Therapeutics (2014), p. 143

[76] B. Haidar, M. Chamoun, A. Serhrouchni, Multilingual cyberbullying detection system, detect- ing cyberbullying in Arabic content, in 1st Cyber Security in Networking Conference (CSNet) (IEEE, 2017)

[77] A. Bellmore, A.J. Calvin, J.-M. Xu, X. Zhu, The five w's of' bullying' on twitter: who, what, why, where, and when, Comput. Human Behav. 44 (March 2015) (2015) 305–314.

[78] C. Salmivalli, Participant role approach to school bullying: implications for intervention, J. Adolesc. 22 (4) (1999) 453–459.

[79] S. Pabian, C.J.S. De Backer, H. Vandebosch, Dark triad personality traits and adolescent cyber-aggression, Pers. Individ. Dif. 75 (March 2015) (2015) 41–46.

[80] S. Alhabash, A.R. McAlister, A. Hagerstrom, E.T. Quilliam, N.J. Rifon, J. Richards, Between likes and shares: effects of emotional appeal and virality on the persuasiveness of anticyberbullying messages on facebook, Cyberpsychol. Behav. Soc. Netw. 16 (3) (2013) 175–182.

[81] V. Balakrishnan, S. Khan, T. Fernandez, H.R. Arabnia, Cyberbullying detection on twitter using big five and dark triad features, Pers. Individ. Dif. 141 (15 April 2019) (2019) 252–257.

[82] M.P. Hamm, A.S. Newton, A. Chisholm, J. Shulhan, A.M. Milne, P. Sundar, H. Ennis, S. Scott, L. Hartling, Prevalence and effect of cyberbullying on children and young people: a scoping review of social media studies, JAMA Pediatr. 169 (8) (2015) 770–777.

[83] A. Ioannou, J. Blackburn, G. Stringhini, E. De Cristofaro, N. Kourtellis, M. Sirivianos, From risk factors to detection and intervention: a practical proposal for future work on cyberbullying, Behav. Inf. Technol. 37 (3) (2018) 258–266.

[84] K. Gahagan, J.M. Vaterlaus, L.R. Frost, College student cyberbullying on social networking sites: conceptualization, prevalence, and perceived bystander responsibility, Comput. Human Behav. 55 (Part B (February 2016) (2016) 1097–1105.

[85] A. Chan, J.S. Antoun, K.C. Morgaine, M. Farella, Accounts of bullying on twitter in relation to dentofacial features and orthodontic treatment, J. Oral Rehabil. 44 (4) (2017) 244–250.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

[86] R. Dredge, J.F.M. Gleeson, X. de la Piedad Garcia, Risk factors associated with impact of severity of cyberbullying victimization: a qualitative study of adolescent online social networking, Cyberpsychol. Behav. Soc. Netw. 17 (5) (2014) 287–291.

[87] M.C. McHugh, S.L. Saperstein, R.S. Gold, Omg u# cyberbully! An exploration of public discourse about cyberbullying on twitter, Health Educ. Behav. 46 (1) (2019) 97–105.

[88] F. Resnik, A. Bellmore, J.-M. Xu, X. Zhu, Celebrities emerge as advocates in tweets about bullying, Transl. Issues Psychol. Sci. 2 (3) (2016) 323–334.

[89] G. Sterner, D. Felmlee, The social networks of cyberbullying on twitter, Int. J. Technoethics 8 (2) (2017) 1–15.

[90] E. Shultz, R. Heilman, K.J. Hart, Cyber-bullying: An exploration of bystander behavior and motivation, Cyberpsychology 8 (4) (2014). Article 3.

[91] A.J. Calvin, A. Bellmore, J.-M. Xu, X. Zhu, #bully: uses of hashtags in posts about bullying on twitter, J. Sch. Violence 14 (1) (2015) 133–153.

[92] R. Dredge, J. Gleeson, X. de la Piedad Garcia, Cyberbullying in social networking sites: an adolescent victim's perspective, Comput. Human Behav. 36 (July 2014) (2014) 13–20.

[93] S.-H. Lee, H.-W. Kim, Why people post benevolent and malicious comments online, Association for Computing Machinery, Commun. ACM 58 (11) (2015) 74–79.

[94] L. Bowler, C. Knobel, E. Mattern, From cyberbullying to well-being: a narrative- based participatory approach to values-oriented design for social media, J. Assoc. Inf. Sci. Technol. 66 (6) (2015) 1274–1293

[95] A. Sengupta, A. Chaudhuri, Are social networking sites a source of online harassment for teens? Evidence from survey data, Child. Youth Serv. Rev. 33 (2) (2011) 284–290.

[96] P.B. Lowry, G.D. Moody, S. Chatterjee, Using it design to prevent cyberbullying, J. Manag. Inf. Syst. 34 (3) (2017) 863–901.

[97] C.M. Kokkinos, E. Baltzidis, D. Xynogala, Prevalence and personality correlates of facebook bullying among university undergraduates, Comput. Human Behav. 55 (Part B (February 2016)) (2016) 840–850.

[98] M. Hood, A.L. Duffy, Understanding the relationship between cyber-victimisation and cyber-bullying on social network sites: the role of moderating factors, Pers. Individ. Dif. 133 (15 October 2018) (2018) 103–108.

[99] G.C.E. Kwan, M.M. Skoric, Facebook bullying: an extension of battles in school, Comput. Human Behav. 29 (1) (2013) 16–25.

[100] A. Lyndon, J. Bonds-Raacke, A.D. Cratty, College students' facebook stalking of ex-partners, Cyberpsychol. Behav. Soc. Netw. 14 (12) (2011) 711–716.

[101] D. Wegge, H. Vandebosch, S. Eggermont, M. Walrave, The strong, the weak, and the unbalanced: the link between tie strength and cyberaggression on a social network site, Soc. Sci. Comput. Rev. 33 (3) (2014) 1–28.

[102] E. Whittaker, R.M. Kowalski, Cyberbullying via social media, J. Sch. Violence 14 (1) (2015) 11–29.

[103] C.J. Case, D.L. King, Internet trolling in social networking sites: a preliminary investigation of undergraduate student victimzation, J. Bus. Behav. Sci. 29 (Fall 2017) (2017) 32–43.

[104] J. Chapin, Adolescents and cyber bullying: the precaution adoption process model, Educ. Inf. Technol. 21 (4) (2016) 719–728.

[105] J. Chapin, Adolescents and cyber bullying: the precaution adoption process model, Educ. Inf. Technol. 21 (4) (2016) 719–728.

[106] S. Horner, Y. Asher, G.D. Fireman, The impact and response to electronic bullying and traditional bullying among adolescents, Comput. Human Behav. 49 (August 2015) (2015) 288–295.

[107] Y. Ophir, C.S.C. Asterhan, B.B. Schwarz, The digital footprints of adolescent depression, social rejection and victimization of bullying on facebook, Comput. Human Behav. 91 (February 2019) (2019) 62–71.

[108] J.V. Peluchette, K. Karl, C. Wood, J. Williams, Cyberbullying victimization: do victims' personality and risky social network behaviors contribute to the problem? Comput. Human Behav. 52 (November 2015) (2015) 424–435

[109] M. Wright, Cyberbullying victimization through social networking sites and adjustment difficulties: the role of parental mediation, J. Assoc. Inf. Syst. 19 (2) (2018) 113–123.

[110] M. Obermaier, N. Fawzi, T. Koch, Bystanding or standing by? How the number of bystanders affects the intention to intervene in cyberbullying, New Media Soc. 18 (8) (2014) 1–7.

[111] S.D. Freis, R.A.R. Gurung, A facebook analysis of helping behavior in online bullying, Psychol. Pop. Media Cult. 2 (1) (2013) 11–19.

[112] S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. DeSmet, I. De Bourdeaudhuij, Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully, Comput. Human Behav. 31 (February 2014) (2014) 259–271.

[113] J. Barlinska, A. Szuster, M. Winiewski, The role of short- and long-term cognitive empathy activation in preventing cyberbystander reinforcing cyberbullying behavior, Cyberpsychol. Behav. Soc. Netw. 18 (4) (2015) 241–244.

[114] B. Cao, W.-Y. Lin, How do victims react to cyberbullying on social networking sites? The influence of previous cyberbullying victimization experiences, Comput. Human Behav. 52 (November 2015) (2015) 458–465.

[115] S. Gearhart, W. Zhang, Gay bullying and online opinion expression: testing spiral of silence in the social media environment, Soc. Sci. Comput. Rev. 32 (1) (2014) 18–36.

[116] B. Holfeld, Perceptions and attributions of bystanders to cyber bullying, Comput. Human Behav. 38 (September 2014) (2014) 1–7.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

[117] T. van Laer, The means to justify the end: combating cyber harassment in social media, J. Bus. Ethics 123 (August 2014) (2014) 85–98.

[118] A.N.-M. Leung, N. Wong, J.M. Farver, You are what you read: the belief systems of cyber-bystanders on social networking sites, Front. Psychol. 9 (365) (2018) 1–11.

[119] H. Machackova, J. Pfetsch, Bystanders' responses to offline bullying and cyberbullying: the role of empathy and normative beliefs about aggression, Scand. J. Psychol. 57 (2) (2016) 169–176.

[120] H.L. Schacter, S. Greenberg, J. Juvonen, Who's to blame?: the effects of victim disclosure on bystander reactions to cyberbullying, Comput. Human Behav. 57 (April 2016) (2016) 115–121.

[121] M.A. Al-garadi, K.D. Varathan, S.D. Ravana, Cybercrime detection in online communications: the experimental case of cyberbullying detection in the twitter network, Comput. Human Behav. 63 (October 2016) (2016) 433–443.

[122] P. Gal´an-García, J.Gdl. Puerta, C.L. Go´mez, I. Santos, P.G. Bringas, Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying, Log. J. IGPL 24 (1) (2016) 42–53.

[123] R. Garett, L.R. Lord, S.D. Young, Associations between social media and cyberbullying: a review of the literature, mHealth. 2 (2016) 46.

[124] T. Milosevic, Social media companies' cyberbullying policies, Int. J. Commun. 10 (2016) 5164–5185.

[125] Mangaonkar, A.; Hayrapetian, A.; Raje, R. Collaborative Detection of Cyberbullying Behavior in Twitter Data. In Proceedings of the Electro/Information technology (EIT), Dekalb, IL, USA, 21–23 May 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 611–616.

[126] Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 1 June 2016; pp. 88–93.

[127] Zhao, R.; Zhou, A.; Mao, K. Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. In Proceedings of the 17th International Conference on Distributed Computing and Networking, New York, NY, USA, 4 January 2016; pp. 1–6.

[128] Singh, V.K.; Huang, Q.; Atrey, P.K. Cyberbullying Detection Using Probabilistic Socio-Textual Information Fusion. In Proceedings of the Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 884–887.

[129] Available online: https://www.Ra.Ethz.Ch/Cdstore/Www2009/Caw2.Barcelonamedia.Org/Index.Html (accessed on 3 July 2020).

[130] Al-garadi, M.A.; Varathan, K.D.; Ravana, S.D. Cybercrime Detection in Online Communications: The Experimental Case of Cyberbullying Detection in the Twitter Network. Comput. Hum. Behav. 2016, 63, 433–443.

[131] Hosseinmardi, H.; Rafiq, R.I.; Han, R.; Lv, Q.; Mishra, S. Prediction of Cyberbullying Incidents in a Media-Based Social Network. In Proceedings of the Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 186–192.

[132] Zhang, X.; Tong, J.; Vishwamitra, N.; Whittaker, E.; Mazer, J.P.; Kowalski, R.; Hu, H.; Luo, F.; Macbeth, J.; Dillon, E. Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network. In Proceedings of the 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 740–745.

[133] Wulczyn, E.; Thain, N.; Dixon, L. Ex Machina: Personal Attacks Seen at Scale. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3 April 2017; pp. 1391–1399.

[134] Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. Int. AAAI Conf. Web Soc. Media 2017, 11, 512–515.

[135] Founta, A.-M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; Kourtellis, N. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In Proceedings of the Weblogs and Social Media (ICWSM), Palo Alto, CA, USA, 25–28 June 2018; pp. 491–500.

[136] de Gibert, O.; Perez, N.; García-Pablos, A.; Cuadros, M. Hate Speech Dataset from a White Supremacy Forum. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium, 31 October 2018; pp. 11–20.

[137] Banerjee, V.; Telavane, J.; Gaikwad, P.; Vartak, P. Detection of Cyberbullying Using Deep Neural Network. In Proceedings of the 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Piscataway, NJ, USA, 15 March 2019; pp. 604–607.

[138] Sadiq, S.; Mehmood, A.; Ullah, S.; Ahmad, M.; Choi, G.S.; On, B.-W. Aggression Detection through Deep Neural Model on Twitter. Futur. Gener. Comput. Syst. 2021, 114, 120–129.

[139] Kumar, A.; Sachdeva, N. A Bi-GRU with Attention and CapsNet Hybrid Model for Cyberbullying Detection on Social Media. World Wide Web 2022, 25, 1537–1550.

[140] Atoum, J.O. Detecting Cyberbullying from Tweets Through Machine Learning Techniques with Sentiment Analysis. In Advances in Information and Communication; Arai, K., Ed.; Springer Nature: Cham, Switzerland, 2023; pp. 25–38.

[141] Hate Speech Twitter Annotations. Available online: https://github.com/ZeerakW/hatespeech (accessed on 9 August 2020).

[142] Wikipedia Detox. Available online: https://github.com/ewulczyn/wiki-detox (accessed on 20 August 2020).

[143] Used a Crowd-Sourced Hate Speech Lexicon to Collect Tweets Containing Hate Speech Keywords. We Use

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8505
***Article Received:** 25 July 2023 **Revised:** 18 September 2023 **Accepted:** 06 October 2023*

Crowd-Sourcing to Label a Sample of These Tweets into Three Categories: Those Containing Hate Speech, Only Offensive Language, and Those with Neither. Available online: https://arxiv.org/abs/1703.04009 (accessed on 16 February 2023).

[144] Hate and Abusive Speech on Twitter. Available online: https://github.com/ENCASEH2020/hatespeech-twitter (accessed on 22 August 2020).

[145] Hate Speech Dataset from a White Supremacist Forum. Available online: https://github.com/Vicomtech/hate-speech-dataset (accessed on 18 November 2022).

[146] Available online: https://www.Kaggle.Com/Datasets/Dataturks/Dataset-for-Detection-of-Cybertrolls (accessed on 10 February 2022).

[147] Lingiardi, V.; Carone, N.; Semeraro, G.; Musto, C.; D'amico, M.; Brena, S. Mapping Twitter Hate Speech towards Social and Sexual Minorities: A Lexicon-Based Approach to Semantic Content Analysis. Behav. Inf. Technol. 2019, 39, 711–721.

[148] Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A.J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science; Springer: Cham, Switzerland, 2020.

[149] Alsharif, M.H.; Kelechi, A.H.; Yahya, K.; Chaudhry, S.A. Machine Learning Algorithms for Smart Data Analysis in Internet of Things Environment: Taxonomies and Research Trends. Symmetry 2020, 12, 88.

[150] Rout, J.K.; Dalmia, A.; Choo, K.-K.R.; Bakshi, S.; Jena, S.K. Revisiting Semi-Supervised Learning for Online Deceptive Review Detection. IEEE Access 2017, 5, 1319–1327.

[151] Li, Z.; Fan, Y.; Jiang, B.; Lei, T.; Liu, W. A Survey on Sentiment Analysis and Opinion Mining for Social Multimedia. Multimed. Tools Appl. 2019, 78, 6939–6967.

[152] Ay Karakus¸, B.; Talo, M.; Hallaç, I˙.R.; Aydin, G. Evaluating Deep Learning Models for Sentiment Classification. Concurr. Comput.Pract. Exp. 2018, 30, 1–14.

[153] Asghar, M.Z.; Khan, A.; Ahmad, S.; Qasim, M.; Khan, I.A. Lexicon-Enhanced Sentiment Analysis Framework Using Rule-Based Classification Scheme. Peer-Rev. Open Access Sci. J. (PLoS ONE) 2017, 12, e0171649.

[154] Khan, F.H.; Qamar, U.; Bashir, S. Lexicon Based Semantic Detection of Sentiments Using Expected Likelihood Estimate Smoothed Odds Ratio. Artif. Intell. Rev. 2017, 48, 113–138.

[155] Ahmed, M.; Chen, Q.; Li, Z. Constructing Domain-Dependent Sentiment Dictionary for Sentiment Analysis. Neural Comput. Appl.2020, 32, 14719–14732.

[156] Nandhini, B.S.; Sheeba, J.I. Cyberbullying Detection and Classification Using Information Retrieval Algorithm. In Proceedings of the International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET), Tamilnadu, India, 15–16 March 2015; pp. 1–5.

[157] Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, Republic and Canton of Geneva, Switzerland, 3–7 April 2017; pp. 759–760.

[158] Park, J.H.; Fung, P. One-Step and Two-Step Classification for Abusive Language Detection on Twitter. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 4 August 2017; pp. 41–45.

[159] Watanabe, H.; Bouazizi, M.; Ohtsuki, T. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. IEEE Access 2018, 6, 13825–13835.

[160] Wang, W.; Huang, J.-t.; Wu, W.; Zhang, J.; Huang, Y.; Li, S.; He, P.; Lyu, M. MTTM: Metamorphic Testing for Textual Content Moderation Software. In Proceedings of the International Conference on Software Engineering (ICSE), Lisbon, Portugal, 14–20 May 2023; pp. 1–13.

[161] Wang, W.; Huang, J.-t.; Wu, W.; Zhang, J.; Huang, Y.; Li, S.; He, P.; Lyu, M. MTTM: Metamorphic Testing for Textual Content Moderation Software. In Proceedings of the International Conference on Software Engineering (ICSE), Lisbon, Portugal, 14–20 May 2023; pp. 1–13.

[162] Roy, P.K.; Tripathy, A.K.; Das, T.K.; Gao, X.-Z. A Framework for Hate Speech Detection Using Deep Convolutional Neural Network. IEEE Access 2020, 8, 204951–204962.

[163] Yadav, A.; Vishwakarma, D.K. Sentiment Analysis Using Deep Learning Architectures: A Review. Artif. Intell. Rev. 2020, 53, 4335–4385.

[164] A. Cuncic, Factors that indulge people in exhibiting offensive behaviour, 2022, https://www.verywellmind.com/the-psychology-of-cyberbullying-5086615.

[165] C. Faucher, W. Cassidy, M. Jackson, Awareness, policy, privacy, and more: Post-secondary students voice their solutions to cyberbullying, Eur. J. Invest. Health Psychol. Educ. 10 (3) (2020) 795–815.

[166] D. CYBERBULLYING, Preventing Cyberbullying. https://www.endcyberbullying. net/preventing-cyberbullying.

[167] S. Saurel, How to stop and prevent cyberbullying in social media, 2019, https://hotinsocialmedia.com/cyberbullying-in-social-media/

[168] S.K. Arora, Cyberbullying laws in India, Int. J. Law Manage. Humanit. 3 (2020) 351.