International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 9, Issue 1 (2023)
https://doi.org/10.31436/ijpcc.v9i1.374

Cyberbullying Conceptualization, Characterization and Detection in Social Media – A Systematic Literature Review

Wai Hong Woo, Hui Na Chua, May Fen Gan*
Department of Computing and Information Systems, Sunway University, Malaysia.

*Corresponding author: ganmayfen@gmail.com
(Received: 9th January 2023; Accepted: 19th January 2023; Published on-line: 28th January 2023

Abstract— Social media has become the primary form of communication wherein users can share intimate moments online through photos, videos, or posts. At a glance, while this greatly improves interconnectivity between people, it also increases the propensity towards unrestricted acts of Cyberbullying, prompting the need for a data-centric detection system. Unfortunately, these sites generate much metadata, which begs the need for complex Machine Learning (ML) classifiers to categorize these acts accurately. Prior studies on the subject matter only target the topics of Conceptualization, Characterization, and Classification of Cyberbullying individually, so this research aims to provide a more holistic understanding of the subject matter in a continuous, synthesized format. This study found that Cyberbullying differs from Traditional Bullying in key areas of Repetition and Intention. Moreover, multimodal feature sets, as opposed to single feature sets, significantly improve ML classifiers' performance. Lastly, the selection of appropriate ML classifiers and performance metrics is context-dependent. The result of this study presents a consolidated view of relevant parties tackling different aspects of an ML-based automated Cyberbullying detection system so that those assigned tasks can approach them strategically.

Keywords— Cyberbullying, systematic literature review, Social Media, Conceptualization, Characterization, Classification, Machine Learning

1 Introduction

One of the main channels for human communication today is none other than through various social media platforms such as Facebook, Douyin, Twitter, and Reddit. Through these platforms, people can keep in touch with their loved ones, network with like-minded individuals across the globe, and share their opinions and ideas online through site postings. However, despite all its advantages, social media, too, has its pitfalls. For example, the issue of cyberbullying, which, quite menacingly, increases the rate of suicide attempts by 8.7% [1], causes anxiety [2], increases the propensity of substance abuse by 2.5% [3], adversely affects academic performance [4], and a host of other issues.

According to [5], Cyberbullying is defined as a belligerent act performed over and over again with time, with intent, and using an electronic communication medium by an individual or a group of individuals towards a victim who is unable to defend himself or herself. Cyberbullying victims, as such, often feel defenceless, unable to do anything to parry the onslaught of negativity and toxicity directed toward them. Furthermore, most cyberbullying occurrences go unnoticed as an implication of the nature, complexity, and sheer volume of the frequency of its happenings [6]. As bullies can also remain anonymous online, they can leverage this anonymity to go under the radar and become more offensive perpetrators [7]. Moreover, cyberbullying, for instance, via text, usually brings more profound and longer-lasting negative effects compared to traditional bullying because victims and bystanders can see traces of the deed repeatedly as it gets propagated online [8]. The gravity of the issue, it's mishandling, and the desire to curb its undesirable effects further amplifies the need to develop more intelligent tools and techniques capable of characterizing and detecting cyberbullying using present social media datasets.

Findings from our literature study show that different detection techniques have been employed to detect cyberbullying in social media. These techniques range from textual analysis methods to more complex and elaborate Machine Learning methods. Although previous literature reviews mainly focused only on the detection process, there is a lack of a study that systematically reviews the conceptualization and characterization processes involved in detecting cyberbullying. Conceptualization, characterization, and detection, when reviewed collectively and sequentially, are complementary and supplementary to a more comprehensive understanding of cyberbullying in the ever-changing technological landscape. Thus, the primary motivation of this systematic literature review is to provide an overview of the current state of affairs concerning the conceptualization, characterization, and detection of cyberbullying. In addition, this review is a

starting point to ascertain gaps in the field of study and incentivize future work and improvements in the three respective disciplines.

This study aims to differentiate the conceptualization of cyberbullying from traditional bullying and compare and contrast different cyberbullying characterization methods and different Machine Learning algorithms for cyberbullying detection. This study tackles the three research questions (RQ) as follows, listed along with the motivation for each RQ:

RQ 1: How is the conceptualization of cyberbullying different from traditional bullying?

Motivation for RQ 1: To identify meta-themes from prior studies contributing to cyberbullying and traditional bullying, and compare their similarities and differences. By distinguishing between the two, feature ambiguity can be significantly reduced, and cyberbullying can be characterized more accurately.

RQ 2: How do we characterize cyberbullying?

Motivation for RQ 2: Identify textual and non-textual characterization methods used for cyberbullying detection in the current literature to establish a more holistic and multimodal approach to cyberbullying characterization. This study also analyses the features used in each characterization method and suggests the rationale for the relevance of the feature in denoting cyberbullying in social media postings.

RQ 3: How can Machine Learning (ML) algorithms be used to detect cyberbullying in social media? and what are the common algorithms used for its detection?

Motivation for RQ 3: Provide a thorough, state-of-the-art look into the process of using ML algorithms to detect cyberbullying in social media through four sequential steps of data collection, feature engineering, ML and evaluation, and in doing so, disseminate the knowledge used in each area to assist in the direction of future research.

2 Related Work

Cyberbullying detection in SM is the title and objective of this study but instead of directly addressing the ML methods used in the detection of cyberbullying, a three-pronged approach is implemented in this study including the conceptualization of cyberbullying, characterization of cyberbullying and detection of cyberbullying, which are highly coupled with each other. This is because better understanding of the definition of cyberbullying leads to better characterization of cyberbullying and better characterization results in improved feature extraction in the eventual detection ML algorithm. Therefore, implementing a three-pronged approach.

A. The Conceptualization of Cyberbullying

The conceptualization of cyberbullying is an essential precursor to the other two elements because having a set of meta-themes to identify and compare traditional bullying or cyber aggression. Common meta-themes or foundational elements indicative of cyberbullying include intent, repetition, accessibility, anonymity, barriers to disclosure, and power imbalance.

The intent is a common recurring meta-theme, especially in how young people conceptualize cyberbullying [9-13]. The intent, in itself, can take varying forms. For example, [14] found that jealousy drives cyberbullying propensity, and [15] reported that vengeance as the intent is a motivating factor. In contrast, [16] emphasized the intent of amusement. Langos [17], on the other hand, discusses how in the absence of intent, accidental behaviours and usual jests are labelled as cyber aggression instead of cyberbullying. Finally, [18] further reinforces the idea, finding that intention is a differentiating factor and a binding agent that verifies an act as cyberbullying. Therefore, all other meta-themes of cyberbullying, despite being present, will be invalidated without intention.

Repetition as a meta-theme of cyberbullying usually indicates some significant intention because cyberbullying is usually not a one-off event. Naruskov et al. [10] and Nocentini et al. [11] recognized that repetition could differentiate between wisecracks and full-blown cyberbullying. SM, unfortunately, promotes this repetition because of its fundamental nature, allowing users to create posts and forward or share the posts of others multiple times, causing these posts to be viewed over and over again. A single-act transition into a repetitive one the more frequently the post, video, or photo is viewed [19].

Accessibility is another symbolic element in cyberbullying. Mishna et al. [20] and Pelfrey Jr and Weber [21] argue that feelings of isolation in younger people increase with the absence of SM. Mishna et al. [20] further added that SM becomes an inviting space for them to openly and unsparingly indulge in derogating their peers. Pelfrey Jr and Weber [21] also concluded that perpetrators of cyberbullying are empowered to continue with their acts upon returning to cyberspace.

Anonymity is another prominent feature and progenitor of cyberbullying. According to [22], the detrimental effects of cyberbullying are compounded by an unknown factor, which increases victims' distress. Similarly, [23] provided examples of different forms of anonymity, such as deception and hacking are prime motivators for cyberbullying as the perpetrators do not need to identify as in real life and can still get away without having to deal with the consequences of their actions.

In the context of barriers to disclosure, [24] argued how fear of repercussions would prevent young people from telling adults that they are victims of cyberbullying. They also found out that this lack of disclosure is because adults are often perceived as incompetent in dealing with matters

on SM. Studies like [14] and [22] also reported how refusal to part with SM or the Internet is another significant barrier to cyberbullying disclosure.

Lastly, power imbalance or discrepancies in capacity between victim and perpetrator is evident in many instances of cyberbullying. Distinct differences in physique, popularity, gender, and intelligence between perpetrators and victims of cyberbullying in favour of the perpetrators can create an imbalance and render the victims powerless in the face of their offenders [25]. Additionally, sometimes a power differential can be made through self-victimization on the victim's part in cases where the victim has poor social interaction skills [26]. Sometimes, the situation is not within their control, as online posting times do not follow a predictable pattern [27].

B. The Characterization of Cyberbullying

Cyberbullying characterization looks into establishing an understanding of the features of cyberbullying for detection. More specifically, what features are extracted, are there any pre-processing requirements, and did the prior study use external resources [28]? Preliminary research into several prior studies for this literature review shows that there are generally four different types of features: content-based, sentiment-based, user-based, and network-based.

1. Cyberbullying Content-Based Features

Content-based features include cyberbullying keywords, profanity, pronouns, n-grams, Bags-of-Words (BoW), Term Frequency Inverse Document Frequency (TFIDF), length of a document, and spelling. Profanity lexicons were created using libraries such as noswearing.com and urbandictionary.com to detect profanity features [29-31]. Rafiq et al. [32], however, were not in favour of using profanity because not all profanities are indicative of cyberbullying, which are just means of expression. Studies such as [29] and [33] used n-grams as the feature of choice in detecting cyberbullying. However, [34] do not favour n-grams, choosing to implement TFIDF instead, claiming that it outperforms n-grams due to its establishment of word importance relative to the document. Sood et al. [35] and Huang et al. [36] are novel studies using less common features, Levenshtein Distance and emoticons, respectively. According to [37], sentiment or emotion-based analysis has applications in product reviews in SM, whereas [38] found that it can be used to dissect patterns in the financial market.

2. Cyberbullying Sentiment-Based Features

Sentiment or emotion-based features, through the identification of keywords, were used in most of the studies screened [39-41]. Nevertheless, this approach is not utilized in [42], which used Probabilistic Latent Semantic Analysis (PLSA) instead to extract the emotive features from the SM postings. In general, sentiment-based features are found to improve the subsequent detection process. However, improvements were insignificant because sentiments can be shrouded in sarcasm, and emotions are not always genuinely expressed, making this feature subpar at best [29, 43, 44].

3. Cyberbullying User-Based Features

Besides content-based and sentiment-based features, user-based features such as age, sex, ethnicity, and sexual preference are also used to detect cyberbullying. Studies like as [45], [46], and [47] used either age or sex or both, with these two features being the most commonly used user-based feature from all the studies screened. Salawu et al. [28] proved that user-based features, like age and sex, substantially improved the detection of cyberbullying trained using ML classifiers.

4. Cyberbullying network-based Features

Lastly, several studies have also reflected the use of network-based features such as the number of likes, friends, followers/following, uploads, and comments. For example, [45] used the network-based feature such as the amount of time spent online, whereas [42], [36], and [48] used ego networks, a type of social network. In particular, [36] found that a higher degree of interconnectedness in an ego network correlates to lower occurrences of cyberbullying, potentially due to better social support. However, more communication between users in that scenario led to more remarkable occurrences of cyberbullying, which is a contrary effect.

Predominantly speaking, earlier works did not limit themselves to only using a single feature type but rather a combination (two or more) of different features for detection. For instance, one of the studies used content-based, sentiment-based, and network-based features to perform detection using ML techniques such as Naïve Bayes, RandomForest, and Decision-Tree [32].

C. The Detection of Cyberbullying

Machine Learning (ML) comprises several but typically four main steps: 1) Data collection, 2) Feature Engineering, 3) Learning, and 4) Evaluation. Data collection is a process by which data is pooled from online databases. A glance into the prior studies screened showed that most of the data were obtained from either SM sites, media sharing platforms, or online databases like Kaggle, with each study dealing with enormous amounts of data, justifying the need for an automated approach in detection. After data is collected, features are extracted in the feature engineering process. According to [49], feature engineering is a process of tweaking the feature space in a dataset to boost the performance of the eventual modelling task. For modelling,

this study focuses on none other than ML. ML approaches can be further categorized into supervised and unsupervised learning approaches.

In terms of supervised learning approaches, the Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers are the two most commonly used techniques for detecting cyberbullying in social media [8, 29, 43, 44, 50]. Studies done using SVM and NB only differs from each other in terms of features. Nandhini and Sheeba [51], a study that used NB, used the Levenshtein Distance, a content-based feature, whereas another NB study, by [46], trained its model using gender, a user-based feature. Even amongst studies that use the SVM approach, the feature engineering process may differ. For instance, [52] relied on keywords on cyberbullying (content-based feature), whereas [35] made use of TFIDF and n-grams (content-based features).

Likewise, between studies that use the NB approach. Notably, [53] used likes and followers/following, a network-based feature, but [54] utilized n-grams, a content-based feature. After SVM and NB, Decision Tree (DT) appeared as another supervised learning approach used, though not as frequent as the two formers, in studies like [32], [47], and [55]. Galán-García et al. [55], in particular, incorporated the J48, a DT classifier. According to [28], DT ensembles such as the J48 are often used because, unlike SVM, they can classify non-linear data without having to map the dataset into a higher dimension. The study also added that the typical reason why SVM and NB are often selected is that both are responsive toward any optimization in their parameters. However, if speed is of concern, NB is preferred over SVM. Preliminary screening into several prior studies did not seem to single out any supervised machine learning techniques in terms of performance because performance ultimately boils down to feature engineering.

The selection of different features under different circumstances will yield different performance results. Thus, in the prior studies screened, there were no supervised ML techniques that consistently came up on top in terms of performance. Besides, there were also no unsupervised ML techniques in the prior studies screened simply because supervised ML techniques were more commonly used. Di Capua et al. [56] and Cheng et al. [57] are two studies favouring the paradigm shift from supervised to unsupervised learning approaches. Di Capua et al. [56] argued that an unsupervised ML approach in the form of an Artificial Neural Network (ANN) could easily outperform its supervised counterparts as it does not need to deal with the complexities of having to label the dataset manually. When dealing with colossal amounts of SM data, this task becomes unfeasible and cannot be generalized for future use cases when the SM evolves.

On the other hand, [57] proposed an unsupervised learning framework called UCD based on time series processing. Therefore, the efficacy of the ML algorithm must be evaluated using several metrics. Commonly used metrics to evaluate the performance of ML algorithms include true positive (TP), true negative (TN), false positive (FP), false negative (FN), recall, specificity, precision, F-score or F-measure, receiver operating characteristic (ROC) curves, area under ROC curves (AUC) and overall accuracy [58, 59].

D. Systematic Literature Review (SLR)

An SLR is a type of literature review that groups all empirical data according to some pre-indicated inclusion criteria to answer a particular set of research questions [60]. An SLR is a meta-analysis with qualities such as having a set of research questions. To which the study will attempt to answer a set of goals that are explicit and easily replicated that meet the eligibility criteria, a quality assessment of the mined literature, a synthesized presentation of data obtained from the selected works, and a high research value [61-63].

3 Research Methodology

We adopted the SLR model for this study to provide a synthesized view of three separate yet interrelated research areas. The SLR protocol in this study has a few stages:

- Identification and Specification of Research Questions
- Sourcing Process
- Inclusion and Exclusion Criteria
- Search and Selection Process
- Quality Assessment
- Data Extraction, Aggregation, and Analysis

For the identification and specification of Research Questions (RQ), RQ1, RQ2, and RQ3, have been detailed in the Introduction section.

A. Sourcing Process

The following are the e-journal databases used in the SLR of this study: IEEE Xplore (https://ieeexplore.ieee.org), Google Scholar (https://scholar.google.com), ScienceDirect (https://www.sciencedirect.com/), and ACM Digital Library (https://dl.acm.org/).

B. Inclusion and Exclusion Criteria

The inclusion and exclusion criteria will differ between the three areas of research in the study: conceptualization, characterization, and detection of cyberbullying in SM. Nonetheless, the general criteria for inclusion and exclusion, are shown in TABLE.

After the first pass of filtering using EC1 to EC6, the pool of available articles to work with is of higher quality and relevance to the study. Therefore, the next step was to

https://doi.org/10.31436/ijpcc.v9i1.374

perform specific filtering using additional exclusion criteria to aggregate further the papers based on the three different RQ areas of conceptualization, characterization, and detection. These criteria are described in Table II, III, and IV.

C. Search and Selection Process (SSP)

The SSP can be divided into several distinct stages, as shown in Fig. 1. Papers excluded from Stage 8 and papers found via backward searching are fed again into Stage 1 of the cycle, and the whole SSP repeats itself, leading to the next iteration.

TABLE I
General Exclusion and Inclusion Criteria

Exclusion Criteria	Inclusion Criteria
EC1: Paper is not written in the English Language	IC1: Paper is written in the English Language.
EC2: Paper is not fully accessible.	IC2: Paper is fully accessible.
EC3: Paper is not peer-reviewed.	IC3: Paper is peer-reviewed.
EC4: Paper is not published in any of the above listed databases.	IC4: Paper is published in one of the above listed databases.
EC5: The content of the paper is not related to the topic and area of study. Paper does not have keywords such as "cyberbullying", "social", "media", "literature", "review", "meta-analysis", "data", "mining", "systematic", "multimodal", "problems", "challenges" or "holistic" in its title.	IC5: The content of the paper is related to the topic and area of study. Paper has keywords such as "cyberbullying", "social", "media", "literature", "review", "meta-analysis", "data", "mining", "systematic", "multimodal", "problems", "challenges" or "holistic" in its title.
EC6: Paper is not published between 2005 and 2021.	IC6: Paper is published between 2005 and 2021.

TABLE II
Exclusion and Inclusion Criteria for RQ1 (Cyberbullying Conceptualization)

Exclusion Criteria	Inclusion Criteria
EC7: Paper does not have keywords such as "conceptualization, "v.s.", "difference", "nature", "motives", "principles", "factors", "meta-themes", "concept", "features" or "themes" in its title.	IC7: Paper has keywords such as "conceptualization", "v.s", "difference", "nature", "motives", "principles", "factors", "meta-themes", "concept", "features" or "themes" in its title.

TABLE III
Exclusion and Inclusion Criteria for RQ2 (Cyberbullying Characterization)

Exclusion Criteria	Inclusion Criteria
EC8: Paper does not have keywords such as "characterization", "analysis", "textual", "sentiment", "lexicon", "NLP", "empirical" or "quantitative" in its title.	IC8: Paper has keywords such as "characterization", "analysis", "textual", "sentiment", "lexicon", "NLP", "empirical" or "quantitative" in its title.

TABLE IV
Exclusion and Inclusion Criteria for RQ3 (Cyberbullying Detection)

Exclusion Criteria	Inclusion Criteria
EC9: Paper does not have keywords such as "detection", "machine", "learning", "supervised", "unsupervised", "comparative", "analysis", "comparison", "models", "training" or "automated" in its title.	IC9: Paper has keywords such as "detection", "machine", "learning", "supervised", "unsupervised" , "comparative", "analysis", "comparison", "models", "training" or "automated" in its title.

D. Quality Assessment (QA)

QA is an integral component of the SSP, performed throughout the SSP lifecycle whenever a set of papers gets screened using any exclusion criteria. QA aims to tackle biases and verify each selected paper's internal and external legitimacy [64, 65]. Selected academic papers for this study were written in English, fully accessible, peer-reviewed, published in one of the electronic databases between 2005 and 2021, and contain all generic and specific keywords for RQ1, RQ2, and RQ3, but that does not necessarily make it a high-quality paper. Therefore, a set of questions [66, 67] were answered for each paper to assess the quality:

- Is the objective of the research adequately described?
- Is the research methodology used in the study sufficiently described?
- Are the results of the studies presented clearly?
- Were the conclusions made by the authors supported by the results of the study?
- Are threats to the legitimacy of the study considered?

E. Data Extraction, Aggregation and Analysis

Data extraction or Coding in this SLR model refers to extracting relevant information from selected papers to answer the three RQs set at the beginning of the study. It can be represented in a tabular format consisting of authors' columns, and year published, objectives, research methodology, tools used, findings, conclusion, limitations, future work, and evidence for RQ1/RQ2/RQ3. The data extraction form for each selected paper was collated into a combined dataset and aggregated based on the three different RQs.

4 Results and Discussion

A. *RQ1: How is the conceptualization of cyberbullying different from traditional bullying?*

Conceptualization of Cyberbullying is paramount to establishing a ground zero on the topic of Cyberbullying.

https://doi.org/10.31436/ijpcc.v9i1.374

Often, the definition is unclear, making the subsequent processes of Feature Engineering, otherwise known as Representation Learning, and then Classification, subpar in its effectiveness, as everything can be traced back to how concrete the conceptualization of Cyberbullying was made. With a clear understanding of the meta-themes that accounts for Cyberbullying, researchers can better understand the corpus they are dealing with, allowing for more accurate labelling by human annotators through some predefined criteria. These predefined criteria make up the six core meta-themes of Cyberbullying. They include themes: 1) ***Intent***, 2) ***Repetition***, 3) ***Power Imbalance***, 4) ***Anonymity***, 5) ***Barriers to Disclosure***, and 6) ***Accessibility***.

On the other hand, Traditional Bullying has three core meta themes which are 1) ***Intent,*** 2) ***Repetition***, and 3) ***Power Imbalance***. We can see that both Cyberbullying and Traditional Bullying share three similar meta themes: ***Intent***, ***Repetition***, and ***Power Imbalance***. The additional factors of ***Anonymity*** and ***Accessibility*** are present in the Cyberbullying conceptualization because they are more specific to the cyber context of SM sites, where these acts of Bullying occur. Although **Barriers to Disclosure** are one of the core meta themes for cyberbullying, it is but a direct consequence of ***Accessibility*** and ***Anonymity***. Consequently, the significant glaring differences between Traditional Bullying and Cyberbullying lie among the three core meta-themes shared, particularly in ***Intent*** and ***Repetition***, with ***Accessibility*** and ***Anonymity*** having less of an impact as much as they are apparent for the context, i.e., occurring in the cyberspace. The prior studies, along with the meta-theme(s) discussed, are shown in Table V.

In increasing order of importance, with the most important being the most frequently discussed in prior studies, the meta themes for cyberbullying can be ranked as follows: Repetition, Intent, Power Imbalance, Anonymity, Accessibility, and Barriers to Disclosure. As expected, the top 3 meta themes for Cyberbullying are all the core meta themes for Traditional Bullying because, dismissing the fact that it occurred in cyberspace, Cyberbullying is still another form of bullying hence the similarity. Barriers to Disclosure are seldom discussed, as seen from the studies in Table VII, most likely because it is related to the reasoning behind why Cyberbullying continues to be rampant rather than being a defining factor for it. On the other hand, for Traditional Bullying, Intent, Repetition, and Power Imbalance were discussed in all relevant papers. All three core meta-themes of Traditional Bullying (Intent, Repetition, and Power Imbalance) are of equal importance as they conceptualize what bullying is in general.

Where applicable, we discuss the commonalities and differences between each meta-theme in the context of both Cyberbullying and Traditional Bullying in sections 4.1.1 – 4.1.6.

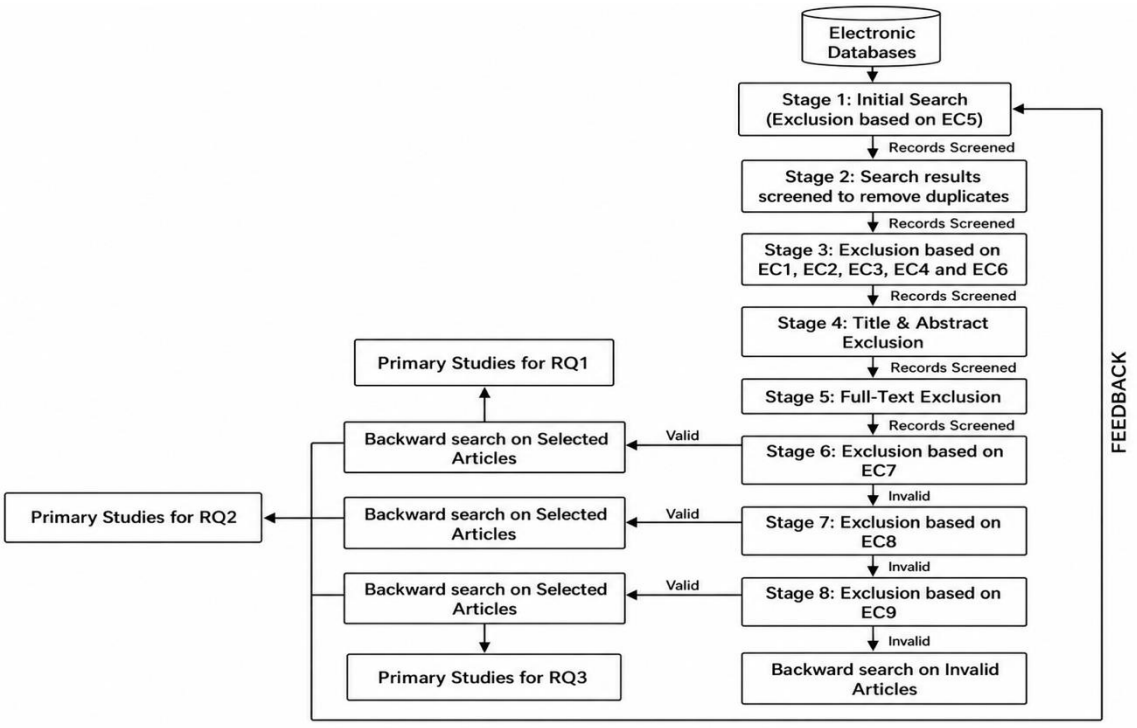

Fig. I SSP Model Diagram

https://doi.org/10.31436/ijpcc.v9i1.374

TABLE V
Core Meta-Themes for Cyberbullying and Traditional Bullying

No.	Study	Intent	Repetition	Power Imbalance	Anonymity	Barriers to Disclosure	Accessibility
Cyberbullying Sources							
1	Slonje and Smith [19]		✓			✓	✓
2	Vandebosch and Van Cleemput [13]		✓	✓	✓		
3	Dooley et al. [27]		✓	✓	✓		✓
4	Mishna et al. [20]						
5	Nocentini et al. [11]	✓	✓	✓	✓		✓
6	Burnham and Wright [24]	✓	✓	✓			✓
7	Hemphill et al. [68]	✓	✓	✓	✓		
8	Langos [17]	✓	✓	✓			
9	Naruskov et al. [10]	✓	✓	✓	✓		✓
10	Ševčíková et al. [69]	✓	✓	✓	✓		
11	Baas et al. [22]	✓	✓		✓	✓	
12	Cassidy et al. [70]	✓	✓	✓	✓	✓	✓
13	Eden et al. [71]	✓	✓	✓	✓	✓	
14	Topcu et al. [12]	✓	✓	✓	✓		✓
15	Berne et al. [9]	✓	✓		✓		
16	Kowalski et al. [72]	✓	✓	✓	✓		✓
17	Pelfrey Jr and Weber [21]	✓	✓	✓	✓	✓	
18	Rafferty and Vander Ven [16]	✓	✓	✓	✓		
19	Thomas et al. [73]	✓	✓	✓	✓		✓
20	Abu Bakar [23]	✓		✓	✓		✓
21	Jacobs et al. [15]	✓	✓	✓	✓		
22	Cuadrado-Gordillo and Fernández-Antelo [74]	✓	✓	✓	✓		✓
23	Fahy et al. [75]		✓	✓			✓
24	Betts and Spenser [14]	✓			✓	✓	✓
25	Dennehy et al. [76]	✓	✓		✓	✓	✓
Traditional Bullying Sources							
1	Aluede et al. [77]	✓	✓	✓			
2	Modecki et al. [78]	✓	✓	✓			
3	Tsaousis [79]	✓	✓	✓			
4	Salin et al. [80]	✓	✓	✓			
5	Stuart and Szeszeran [81]	✓	✓	✓			

1. INTENT

Intent can come in a myriad of forms, but in the case of Cyberbullying and Traditional Bullying, the perpetrator's intentions are usually bad in most cases. According to [9] and [17], bad intentions in Cyberbullying include actions such as making spiteful remarks about a person's looks, sexual preference, and circle of friends. However, bad intentions can also include making sarcastic or snide comments whose reaction can go both ways; either the person receiving said remarks get offended or laugh along with the joke. Therefore, Intent cannot be a standalone criterion for both Cyberbullying and Traditional Bullying, so it has to be used alongside Power Imbalance and Repetition in conceptualizing both forms of bullying [17]. Nocentini et al. [11] further added how the notion of Intent could be individualistic and not so clear-cut in Cyberbullying because the perpetrator may regard it as a harmless prank or joke. However, the victim was offended. Thus, intent as a meta-theme of Cyberbullying should focus on the perceived

intention of the perpetrator to cause harm, causing the victim to feel attacked and unsafe. In Traditional Bullying, the intent of the bully to cause harm is always concretely established [77-81].

2. Repetition

According to [9], SM sites can become a venue for perpetrators, friends of perpetrators, enablers, and mindless bystanders to disseminate information repeatedly. [11] also concurred with the study, adding that there is an element of Repetition in Cyberbullying through the persistent forwarding posts on SM sites. Cyberbullying detection aims to create a failsafe mechanism that, in an ideal world, will detect instances of Cyberbullying at the earliest so that these posts can either be changed or deleted to keep exposure to audiences of the Internet at a minimum. However, according to [17], this could be hard to achieve as most of the time, Cyberbullying posts tend to remain in cyberspace eternally.

Another thing to note is that Repetition in Cyberbullying is often not characterized by actions performed repeatedly by the malefactor. However, it is usually the case where past deeds may have an incessant effect [22]. In Traditional Bullying, bullying is always set in stone when there is an element of repetitiveness inflicted by the perpetrator toward the victim, which consolidates the victim's perception of the perpetrator's intent to cause harm [77-81].

3. Power Imbalance

Power Imbalance exists when the victim is in a vulnerable position, unable to stand up for himself or herself during the perpetrator's display of power toward the victim [17]. In that same vein, even though the victim has perceived differences in power, if he or she is not rendered defenseless, then a power differential does not exist. Langos [17] further added that stark differences in age, gender, physique, mental capacity, sociability, and social standings lead to disparities in power between aggressor and victim in most cases studied. The factors above are valid for both Cyberbullying and Traditional Bullying. Nocentini et al. [11], however, took a different approach in their studies, researching whether victims of Cyberbullying can quickly end the negative interactions between them and the perpetrators, thereby reducing or eliminating any power imbalances. Despite their efforts, the study concluded that Power Imbalance still exists in Cyberbullying, as posts on SM sites can propagate like wildfire, leading to more exposure and deepening of the issue. In Traditional Bullying, Power Imbalance exists as well and is a crucial feature in its conceptualization [77-81].

4. Anonymity

An easy way to comply with the journal paper formatting requirements is to use this document as a template and simply type your text into it.

According to [9], the probability that an anonymous person becomes a victim of Cyberbullying is equal regardless of how authoritative he or she is deemed in real life. This is closely related to identity forgery, which, as discussed in [10], enables people to do the unthinkable and act in ways they would never do in reality. The term Anonymity can also literally refer to the notion that the perpetrator of Cyberbullying commits derogatory acts while on the keyboard, unbeknownst to the victims [20]. Cuadrado-Gordillo and Fernández-Antelo [74] further added that this Anonymity could serve as a "safe house" for the perpetrators, which only works in their favor, making it harder for authorities to rat them out as aggressors. Moreover, this "safe" setting made possible through Anonymity gives perpetrators the impression that they have the authority to hound others without repercussions [20]. Furthermore, this can make it seem like the acts of Cyberbullying are non-orchestrated. However, this is not the case for Traditional Bullying, as the perpetrator's identity is always known [13].

5. Barriers to Disclosure

Barriers to Disclosure, in most cases, are a direct consequence of both Accessibility and Anonymity, or it pertains more to why Cyberbullying is still at large rather than being a defining cause of it. Out of all the studies screened, only two, [22] and [20], stood out as the two with the most explanatory takes on Barriers to Disclosure. According to [22], Barriers to Disclosure take the form of the victim's reluctance to seek assistance. Mishna et al. [20] touched on several Barriers to Disclosure, including victims of Cyberbullying are hesitant to report their bullying to parents and other adult figures because it is a problem that is incomprehensible to them due to generational differences. Furthermore, victims of Cyberbullying are afraid to disclose any information concerning their bullying for fear of having their computer privileges stripped from them. There was also a point made on how victims of Cyberbullying do not bother with telling the truth as the perpetrators will most likely be disingenuous, leveraging on the power of Anonymity. Finally, the study also brought up how most victims of Cyberbullying take a non-disclosure disposition when it comes to experiences such as Cyberbullying in order to be self-sufficient so as not to worry their folks.

6. ACCESSIBILITY

An easy way to comply with the journal paper formatting requirements is to use this document as a template and simply type your text into it.

Accessibility, like Anonymity, is also one of the meta themes specific to the cyber context. However, unlike Anonymity, Accessibility can be quite multi-faceted, referring to several aspects that can explain why Cyberbullying has occurred. The first aspect of Accessibility is related to the ease of access by which perpetrators can commit their heinous acts of Cyberbullying. According to [14], Accessibility refers to the absence of a cut-off time for Cyberbullying, which perpetrators of Cyberbullying can leverage. Traditional bullying often occurs in a physical setting with different cut-off times, such as when one is asleep or engaged in other activities, and facilities restrictions, such as closing hours and barring people from being around the physical compound, thus halting the act of Traditional Bullying. Unlike Cyberbullying, acts can go on around the clock in cyberspace. The second aspect of Accessibility is related to the ease of publicizing the acts of Cyberbullying, forwarding the same post a hundred to a thousand times over through SM, garnering huge volumes of bystanders who may or may not aggravate the situation [10, 74].

B. ***RQ2***: *HOW DO WE CHARACTERIZE CYBERBULLYING*?

In the lens of an automated Cyberbullying detection framework, the Characterization of Cyberbullying is essentially the process of Feature Extraction or Feature Engineering in the landscape of the Cyberbullying detection framework. It is the crucial step that precedes the Classification of Cyberbullying in detection. Feature Extraction converts the data corpus into a format that can be fed into the various Machine Learning algorithms. In addition, various features can be extracted from the corpus. However, features used in prior studies typically fall into one or more of four types of features: 1) content-based features, 2) sentiment/emotion-based features, 3) user-based features, and 4) network-based features.

Content-based features is an umbrella term for features extracted from the content of the post or corpus, such as pronouns, n-grams, Bag-of-Words (BoW), and Term Frequency-Inverse Document Frequency (TF-IDF). Sentiment/emotion-based features, on the other hand, include features such as Cyberbullying-specific keywords, Polarity, and outputs of any Semantic Analysis. In addition, features specific to users, like age, gender, and sexual orientation, make up what is known as user-based features. Finally, network-based features or social network features encompass (but are not limited to) the follower count, the following count, and the number of likes. As content-based features make up anything that can be extracted from the content of SM sites, they can include both textual/lexicon and contextual/non-textual features. For example, boW (Bag-of-Words) is a textual, content-based feature, whereas an audio file is a contextual, content-based feature. This distinction is crucial in understanding the differences between the two terms. Thus, contextual features include various feature types: network-based, sentiment/emotion-based, user-based, and all image and visual features.

According to [36], a multimodal approach must be taken regarding what features are used to train Classification algorithms because the efficacy of using textual features alone is relatively poor. In that study, the contextual feature used in tandem with the traditional textual features is social network features (or network-based features). This multimodal approach led to marked improvements in the classification accuracy of the classifiers employed in the study. Furthermore, Cyberbullying is considered a social issue, so information on the social context encircling the textual information may offer critical insights which can help improve its detection. In layperson's terms, having social network features paints a complete picture in prior studies on Cyberbullying detection, so its inclusion was justified. Rezvani et al. [82] also, second this, claiming to implement a multimodal or combined feature set. Their research on the inclusion of image metadata features to traditional textual features yielded similar improvements. Prior studies relevant to the features used are detailed in Table VI.

TABLE VI
TYPES OF FEATURES USED IN LITERATURE

STUDY	TYPE OF FEATURE			
	CONTENT	SENTIMENT/ EMOTION	USER	NETWORK
BAYZICK ET AL. [83]	INSULT WORDS, SWEAR WORDS, 2ND PERSON PRONOUNS CAPITALIZATION	-NIL-	-NIL-	-NIL-
REYNOLDS ET AL. [84]	NUM (# BAD WORDS), NORM (DENSITY OF BAD WORDS), SUM (OVERALL BAD-NESS OF POST), TOTAL (TOTAL #WORDS IN A POST)	-NIL-	-NIL-	-NIL-

https://doi.org/10.31436/ijpcc.v9i1.374

Dadvar et al. [8]	profane words, 2nd person pronouns, other pronouns, Term Frequency – Inverse Document Frequency (TF-IDF)	-Nil-	gender	-Nil-
Dadvar and De Jong [46]	profane words, 2nd person pronouns, other pronouns, TF-IDF	-Nil-	gender	-Nil-
Dinakar et al. [85]	Profanity, TF-IDF, negativity (Ortony Lexicon), subtlety (Label Specific Features)	-Nil-	-Nil-	-Nil-
Dadvar et al. [86]	profane words, normalized 1st & 2nd person pronouns, profanity window, # emoticons, # cyberbullying words, ratio of capital letters in a comment	-Nil-	user's activity history, age	-Nil-
Kontostathis et al. [87]	Bag of Words (BoW), TF-IDF	-Nil-	-Nil-	-Nil-
Nahar et al. [88]	binary representation of keywords, normalized value of keywords, pronouns, normalized value of capital letters, messages' metadata	normalized values of emotions (happy & angry)	Age, gender	location
Hosseinmardi et al. [53]	Profanity, linguistic, image, cyberaggression, # wordspsychological measurements	-Nil-	-Nil-	# likes, # comments, # following, # followers
Ptaszynski et al. [89]	disjoint elements	-Nil-	-Nil-	-Nil-
Squicciarini et al. [47]	post length, offensive words, 2nd person pronouns	post sentiment (measured using Semantria)	Gender, age, on-site activity	degree centrality, closeness centrality, between-ness centrality, eigenvector centrality, clustering coefficient
Van Hee et al. [90]	BoW	Polarity	-Nil-	-Nil-
Van Hee et al. [91]	BoW	Polarity	-Nil-	-Nil-
Al-Garadi et al. [92]	Vulgarity, words, acronyms and abbreviations of cyberbullying-specific words, 1st & 2nd person pronouns	-Nil-	Personality, gender, age	# followers, # following, following-followers ratio, verified status, # tweets, mentioned users, favourites
Zhao et al. [93]	BoW, word embeddings	Latent Semantic Feature	-Nil-	-Nil-
Escalante et al. [94]	n-grams, TF-IDF	-Nil-	Profile Specific Representation (PSR), Subprofile Specific Representation (SSR)	-Nil-
Haidar et al. [95]	tweet content, language (English & Arabic), cyberbullying keywords	-Nil-	-Nil-	-Nil-
Özel et al. [96]	BoW, TF-IDF, emoticons	-Nil-	-Nil-	-Nil-
Raisi and Huang [97]	bullying bigrams	-Nil-	-Nil-	-Nil-
Singh et al. [98]	informal language, sexual words, 3rd person singular pronouns, tone, psychological words, words that show personal concern, image category, image type, sexual content	-Nil-	Age, gender	-Nil-
Rosa et al. [99]	BoW, TF-IDF, Inverse Class Frequency (ICF)	-Nil-	-Nil-	-Nil-
Hani et al. [100]	n-grams, TF-IDF	Polarity	-Nil-	-Nil-
Balakrishnan et al. [101]	# characters, uppercase characters, lowercase characters, hashtags, symbols, user mentions, URLs, media	sentiment analysis, emotion analysis	verified status, status count, list count, user favourite count,	# followers, # following, popularity (followers-following ratio)

			ACCOUNT AGE, PERSONALITY	
GENCOGLU [102]	SENTENCE EMBEDDINGS	-NIL-	-NIL-	-NIL-
KARGUTKAR AND CHITRE [103]	WORD-TO-VECTOR REPRESENTATION (WORD2VEC)	-NIL-	-NIL-	-NIL-
MUNEER AND FATI [104]	TF-IDF, WORD2VEC	-NIL-	-NIL-	-NIL-
REZVANI ET AL. [82]	TEXTUAL INFORMATION, IMAGE FEATURES	ENRICHMENT FEATURES	-NIL-	# FOLLOWERS, # FOLLOWING, # LIKES, POPULAR CATEGORIES, AVERAGE REACTIONS, AVERAGE REPLIES, # MENTIONS
ATES ET AL. [105]	BOW, TF-IDF	-NIL-	-NIL-	-NIL-
BOZYIĞIT ET AL. [106]	BOW, TF-IDF, OTHER TEXTUAL FEATURES	-NIL-	-NIL-	# RETWEETS, # FAVOURITE, # HASHTAGS, # MENTIONS, ACCOUNT AGE, # FOLLOWERS, # FOLLOWING, # TWEETS, LOCATION, IS SELF-MENTIONED, # LIKES, # MEDIA
PERERA AND FERNANDO [107]	TF-IDF, PROFANITY + PRONOUN, FREQUENCY OF CYBERBULLYING, THEMES/CATEGORIES	POLARITY	-NIL-	-NIL-

*** # denote "Total Number (No.) of" ***

*** -Nil- implies "Not in Line" ***

It was found that ***content-based features*** were adopted in all papers gathered, ***sentiment/emotion-based*** and **user-based features** were 30%, and ***network-based features*** were 23.3% of total papers gathered for Cyberbullying detection. The evolution graphs in Fig. 2, 3, and 4 show how different features used in the Characterization or Feature Engineering process have evolved over the years. Fig. 2 depicts the evolution in the usage of the four types of features over the years. In contrast, Fig. 3 portrays the evolution in the usage of textual and non-textual features over the years. Finally, Fig. 4 displays the evolution of using single and multimodal feature sets over the years. Based on the graph, we observe an apparent tendency that the multimodal features approach is likely to be the direction of future studies for Cyberbullying classification.

As observed in Fig. 3, except for the 2018 study, [99] see the inclusion of contextual/non-textual features in the feature sets used to train the classifiers. Contextual/non-textual features are never used alone but are always used in tandem with textual features. However, in many studies, textual features are used in isolation which is traditionally how most studies were in the earlier times before they evolved to include newer features, both contextual/non-textual features and reengineered textual features in their feature sets. Nonetheless, incorporating contextual/non-textual features improves the ensuing Classification performance [46, 82, 101].

As seen in Fig. 4, except in the 2018 study, [99] evaluated the efficacy of a combined or multimodal feature set, in which case a multimodal/combined feature set is pitted against other multimodal/combined or single feature sets. Cyberbullying detection using ML techniques generally falls into either one of three categories: 1) evaluate the effectiveness of a new/proposed feature by comparing different feature sets, i.e., baseline vs. [baseline + new/proposed feature]. 2) assess the effectiveness of a new/proposed Cyberbullying detection framework/classifier by comparing different classifiers. 3) compare the performance of different classifiers. We discuss the first category in this section., then the second and third categories in the following section on RQ3.

A multimodal/combined feature set is a union between textual and contextual/non-textual features in comparing different feature sets. According to [91], the best Classification performance was achieved with a multimodal/combined approach, merging all the different features and not when the features are used separately or in smaller combinations. According to [82], adding contextual/non-textual features to textual features yielded better results than when using contextual/non-textual features alone or textual features alone, wherein only subpar results were obtained for both cases. Another study by [98] showed that when used on their own, textual features outperformed contextual/non-textual features, contrary to the initial expectations of the researchers. However, the best performance across all classifiers is still observed only with a combined usage of the textual and contextual/non-textual features.

https://doi.org/10.31436/ijpcc.v9i1.374

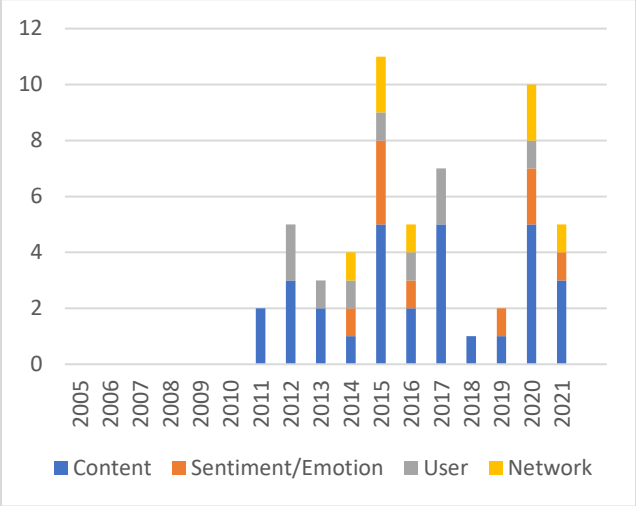

Fig. 2 Evolution in Usage of the 4 Different Types of Features

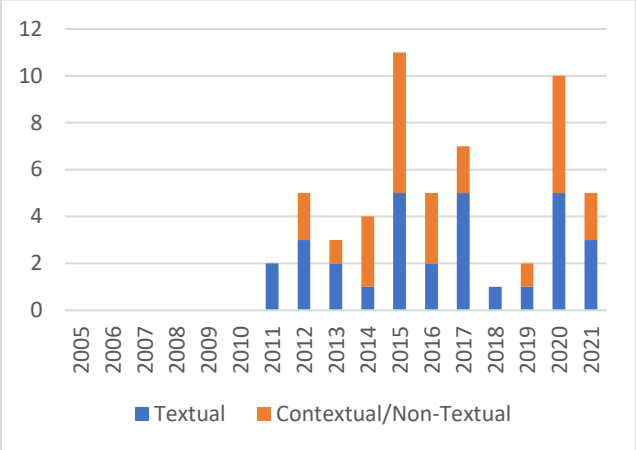

Fig. 3 Evolution in Usage of Textual vs Contextual/Non-Textual Features

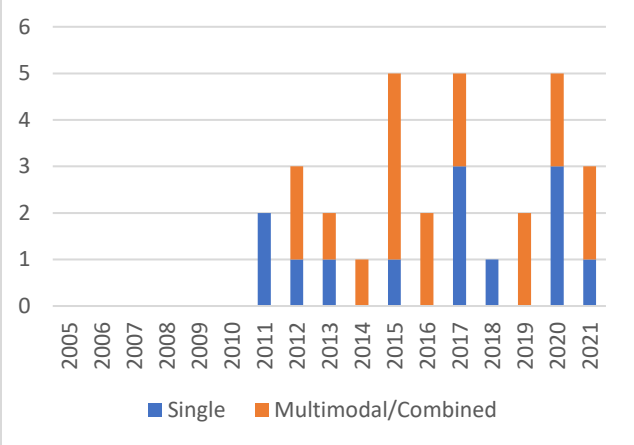

Fig. 4 Evolution in Usage of Single vs Multimodal/Combined Feature Set

It is observed that the combination of contextual/non-textual features and textual features can also be a novel feature on its own, not necessarily a multimodal/combined feature set. For example, in [94], two inventive features called PSR and SSR are by-products of the amalgamation between textual and contextual/non-textual features, which remarkably improved the performance of the classifiers across all measures.

With the findings above, one may tempt to infer: 1) all classifiers unanimously well on a particular feature set, especially the multimodal/combined feature set; 2) adding just about any contextual/non-textual features to current baseline textual features would significantly improve classification performance. Nevertheless, several studies rebuked the previously mentioned inferences. For example, in [100], it was observed that a classifier might perform well on a particular feature set but not on other sets. This finding shows that while it may appear that all classifiers, without opposition, would perform well on one feature set, usually a multimodal/combined feature set; however, it cannot be generalized. So, it is worth conducting a study to ascertain and not make assumptions based on another study. Regarding the second inference on guaranteed improvements with the addition of contextual/non-textual features as part of the multimodal/combined feature set, [53] demonstrated that incorporating such features led to non-statistically significant improvements in the performance of the classifiers. On the other hand, [101] dispelled the second claim by showing that only some contextual/non-textual features give meaningful boosts to the classifiers' performances. This study brings up a need for some way to select only important features from a pool of several features, otherwise known as Feature Selection.

In summary, the features used to train the classifiers responsible for the Cyberbullying prediction in SM postings to fall into either textual or contextual/non-textual instances or, in some fringe cases, a combination of both. The justification for creating these instances is always related to the combined process of conceptualization and characterization of Cyberbullying, i.e., what features can we characterize using the available source of information to help understand, denote or give better context to Cyberbullying? Using textual and contextual/non-textual features results in a multimodal/combined feature set; broadly speaking, ML classifiers tend to perform the best using it as opposed to single feature sets. That said, one should always run an experiment to verify the facts because results may vary from the norm depending on the corpus and classifier used.

https://doi.org/10.31436/ijpcc.v9i1.374

C. *RQ3*: How Machine Learning (ML) algorithms can be used to detect cyberbullying in social media, what are the common algorithms used and which algorithm(s) is the most effective for the purpose of its detection?

The prior two sections, 4.1 and 4.2, lay the grounds for understanding the prerequisites for an ML-based automated Cyberbullying detection system. An ML-based automated Cyberbullying detection system is bipartite [93]. The first component is Representation Learning (used interchangeably with Feature Engineering), including the Conceptualization and Characterization of Cyberbullying (or Feature Extraction). The second component is Classification. Classification aims to produce a system adept at categorizing Cyberbullying given a corpus or SM content and generating results of Classification or other statistical measures. The ML-based classifiers, which are the tools for Classification, fall into either one of two categories: 1) Traditional ML classifiers and 2) Deep Neural Network (DNN) or Deep Learning (DL) approaches. Traditional ML classifiers include Naïve Bayes, Support Vector Machine, and Logistic Regression. In contrast, DNN/DL approaches include Convolutional Neural Networks, Multilayer Perceptron, Kohonen, or Self Organizing Map.

Where applicable, Table VII below shows the relevant prior studies with points of contention including types of features used (Content-based (Cb), User-based (Ub), Network-based (Nb), and Sentiment-based (Sb)), type of classifiers used (ML or DL or NLP or Mixed), and their classification algorithm(s).

Regarding the feature types used, we observe that all relevant prior studies (40 papers) collected used content-based features for automatic Cyberbullying detection, followed by the network-based with 30% of the total papers, 25% for user-based and sentiment-based features.

Furthermore, we found that of the papers collected on automated Cyberbullying detection using ML classifiers, 65% of them used pure ML classifiers. However, none used purely the DL/DNN approaches. For the mixed classifiers approach, we encountered 35% of studies. In the matter of Classification algorithms used, we summarize the overall adoption rates (among all relevant 40 papers gathered) by ML and DL approaches in Table VIII.

TABLE VII
Automatic Cyberbullying Detection in Social Media

Study	Data source	Features type	ML/DL	Classification algorithms
Go et al. [108]	Twitter	Cb	ML	NB, MaxEnt, SVM
Reynolds et al. [84]	Formspring	Cb	ML	J48, JRip, IBK, SMO
Dadvar and De Jong [46]	MySpace	Cb, Ub	ML	SVM
Dadvar et al. [86]	YouTube	Cb, Ub	ML	SVM
Mangaonkar et al. [54]	Twitter	Cb	ML	NB, SVM, LR
Nandhini and Sheeba [51]	Formspring, MySpace	Cb	ML	NB
Al-Garadi et al. [92]	Twitter	Cb, Nb, Ub	ML	NB, SVM, RF, K-NN
Di Capua et al. [56]	Formspring, YouTube, Twitter	Cb, Sb, Nb	Mixed	GHSOM
Singh et al. [109]	Twitter	Cb, Nb	ML	ZeroR, Naïve/Early Fusion, Late Fusion
Sintaha et al. [110]	Twitter	Cb, Sb, Nb, Ub	Mixed	NB, SVC (RBF kernel), SVC (Linear kernel), LinearSVC, CNN
Zhao et al. [93]	Twitter	Cb, Sb	ML	Linear SVM
Romsaiyud et al. [111]	Perverted-Justice, Twitter	Cb, Nb, Ub	ML	NB
Singh et al. [98]	Instagram	Cb	ML	Bagging Classifier
Agrawal and Awekar [112]	Formspring, Twitter, Wikipedia	Cb	Mixed	LR, SVM, RF, NB, CNN, LSTM, BLSTM
Nurrahmi and Nurjanah [113]	Twitter	Cb	ML	SVM, K-NN
Soni and Singh [114]	Vine	Cb, Sb	ML	K-NN, SVM, NB, LR, RF
Tahmasbi and Rastegari [115]	Twitter	Cb, Nb, Ub	ML	J48, JRip, RF, LR, AdaBoost, SVM, NB
Banerjee et al. [116]	Twitter	Cb	Mixed	CNN
Biesek [117]	Twitter	Cb	Mixed	SVM, Bidirectional GRU, Flair Framework (Bidirectional GRU + embeddings)
Cheng et al. [118]	Instagram, Vine	Cb, Nb, Ub	ML	RF, Linear SVM, LR

https://doi.org/10.31436/ijpcc.v9i1.374

Cornel et al. [119]	Ragnarok, DotA	Cb	Mixed	CNN, NB
Haidar et al. [120]	Twitter	Cb	ML	Stacking Ensemble (Simple LR, K-NN, RF, SVM, Bayesian LR, SGD), Boosting & Bagging Ensemble (NB, SVM, K-NN)
Hani et al. [100]	Kaggle	Cb, Sb	Mixed	SVM, NN
Kumar et al. [121]	YouTube	Cb, Sb	ML	RF, SMO, K-NN, NB
Liu et al. [122]	MySpace, Twitter, Facebook, Formspring, Ask.fm, Instagram, Vine, 2 unspecified sources	Cb, Nb, Sb	ML	NB, DT, RF, Tree Ensemble, LR, SVM
Yao et al. [123]	Instagram	Cb	ML	RDFS Classifier, RF, Dynamic LR, TM, CONcISE Classifier
Zhang et al. [124]	Twitter	Cb, Sb, Nb	Mixed	Linear SVM, LR, DT, RF, Gradient Boosting Regression Tree, MLP
Alasadi et al. [125]	Instagram	Cb	ML	NB, Baseline classifier (Bayesian Fusion Model with $\lambda = 0$), Bayesian Fusion Model
Ali and Syed [126]	Twitter, Formspring	Cb, Sb	ML	RF, NB, SVM, LR, Ensemble Classifier
Balakrishnan et al. [101]	Twitter	Cb, Ub, Nb, Sb	ML	RF, J48
Islam et al. [127]	Facebook, Twitter	Cb	ML	DT, NB, RF, SVM
Kumar and Sachdeva [128]	Twitter, Facebook	Cb	Mixed	MIIL-DNN, NB, DT, MLP, SVM, CNN, LSTM
Van Bruwaene et al. [129]	Instagram, Facebook, Pinterest, Twitter, Gmail, YouTube, Tumblr, other unspecified resources	Cb	Mixed	ZeroR, SVM, XGBoost, CNN
Wang et al. [130]	Instagram, Vine	Cb	Mixed	SVM, NB, LR, RF, LSTM, Text-CNN, HAN, MMCD Framework (BiLSTM + HAN + other embeddings)
Ahmed et al. [131]	Facebook	Cb	Mixed	DNN, Ensemble Classifier
Alsubait and Alfageh [132]	YouTube	Cb	ML	Multinomial NB, Complement NB, LR
Azeez et al. [133]	Twitter	Cb, Nb, Ub	ML	Multinomial NB, K-NN, LR, DT, RF, LinearSVC, AdaBoost, SGD, Bagging Classifier, Ensemble Classifier
Eronen et al. [134]	Formspring (English), Dataset from Ptaszynski et al. (2010) (Japanese), Twitter (Polish)	Cb	Mixed	LR, CGD LR, SGD SVM, Linear SVM, K-NN, NB, RF, AdaBoost, XGBoost, MLP, CNN, LBFGS LR, Newton LR
Ge et al. [135]	Instagram, Vine	Cb, Nb, Ub	Mixed	LR, SVM, XGBoost, CNN, LSTM, SelfAtt, HAN, SICD, Soni & Sign (temporal model), HANCD, TGBully (temporal model)
Ghosh et al. [136]	various unspecified sources	Cb	ML	SVM, LR, RF, PA
Chong et al. [137]	various unspecified sources	NLP features	ML	Rule-based and Zero-shot classifier

The current state of the art shows that traditional ML classifiers are more popular than DNN/DL approaches. Lastly, for the Classification model evaluation metrics used, we summarize the usage rate as follows: Accuracy 67.5% (of 41 total papers gathered, applied to the rest of this list), F/F1-measure/score 67.5%, Recall 62.5%, Precision 60%, Area Under Receiver Operating Characteristic Curve (AUC) 27.5%, True Positive (TP) 12.5%, Confusion Matrix: 7.5%, Error 7.5%, True Negative (TN) 5%, False Positive (FP) 5%, False Negative (FN) 2.5%, and other evaluation metrics: 10%.

TABLE VIII
ADOPTION RATE OF CLASSIFICATION ALGORITHMS USED IN PRIOR STUDIES

ML approaches	DL approaches
Support Vector Machine (SVM): 60% Naïve Bayes (NB): 50% Logistic Regression (LR): 37.5% Random Forest (RF):35% K-nearest Neighbours (K-NN):15% Ensemble techniques: 15% Boosting techniques: 15% Decision Tree (DT) :12.5% J48 classifier 7.5% Sequential Minimal Optimization (SMO) :5% ZeroR classifier: 5% JRip classifier: 5% Stochastic Gradient Descent (SGD): 5% Zero-shot classifier: <0.1% Other ML techniques/frameworks: 20%.	Neural Network (NN): 27.5% Long Short-Term Memory (LSTM): 10% Multilayer Perceptron (MLP): 7.5% Attention Network (AN): 5% Self-Organizing Maps (SOM): 2.5% Other DL approaches: 10%

Most of the studies' dataset is manually labelled by either field domain experts or non-professionals from crowdsourcing platforms of Amazon's Mechanical Turk and CrowdFlower [46, 56, 84, 86, 93, 112, 113, 120-125, 129, 131, 134, 136]. Novel labelling techniques also used automated code observed [110, 133] and a score-based system [119]. However, labelling is not always necessary, such as in [101], whose authors used an annotated dataset from a previous study. Romsaiyud et al. [111] worked with both annotated and non-annotated variants of the dataset and concluded that the labelling of datasets led to higher Classification accuracy.

We notice that features extracted fall into one of four broad categories: content-based features, network-based features, sentiment-based features, and user-based features. Moreover, using multiple features together tends to give the best performance compared to using them in isolation [46, 86, 101, 108, 114, 122, 124]. Different features are selected based on the type of classifier used in the studies, even though its underlying type may be the same. For example, the textual content-based features commonly used are Natural Language Processing (NLP) features such as n-gram, BoW, and TF-IDF for traditional ML. For DNN/DL classifiers, it is customary to use textual content-based features in the form of word embeddings such as Word2Vec, Doc2Vec, and GloVe [110]. In [115], an unorthodox approach called Linguistic Inquiry Word Count (LIWC) was used to extract features to avoid a skewed perception that profanity is a defining characteristic of Cyberbullying.

Work in [137] adopted the Zero-shot topic classifier and rule-based methods using Natural Language Processing (NLP) for comparing the accuracy of classifying cyberbullying. The result in [137] shows that the information extraction rule-based model applied provides limited assistance in categorizing cyberbullying behaviours. The Zero-shot model presented a better rate of recognizing the flaming behaviour compared to rule-based modelling, but its accuracy rate dropped when identifying the other cyberbully behaviours. Nevertheless, the finding of [137] concluded that Zero-shot text classification model is useful for predicting flaming behaviour.

Additionally, most studies suggest that non-textual and textual features used in a Multimodal/Combined feature set seem to be the trend. The addition of such features led to statistically significant enhancements in the performance of ML classifiers [46, 86, 98, 101, 108, 114, 122, 124, 130]. When it comes to features, however, adding more and more in combination may not always be suitable for the given context, so sometimes less is more. To [123], the approach proposed can achieve high Classification accuracy with the bare minimum number of features, making the model ideal for large-scale datasets. Strangely as well, some features work better in one language but not as good in other languages, so besides context dependency, there is also some level of language (that of the dataset(s) used) dependency when it comes to features used in a study [134].

5 IMPLICATIONS FOR PRACTICE

Regarding the Conceptualization of Cyberbullying, we saw the delineation of the differences between Traditional Bullying and Cyberbullying through the lens of various meta-themes such as Intent, Repetition, Power Imbalance, Accessibility, Anonymity, and Barriers to Disclosure. Through a simple one-on-one comparison between the two, it is clear that it would be folly to assume that Cyberbullying is the online equivalent of Traditional Bullying due to the differences in conceptualizations between the two. That said, researchers looking to improve the initial annotation process of Cyberbullying/Non-Cyberbullying should be looking into key factors such as Repetition and Intent in

deciphering whether the text is Cyberbullying or Non-Cyberbullying. Non-Cyberbullying is a blanketing, all-inclusive term that can include posts containing harmless pranks, satire, or sarcasm to posts that are non-bullying altogether. Traditional Bullying is a subset of Non-Cyberbullying, so the distinction between Cyberbullying and Traditional Bullying is crucial to the success of the forthcoming ML procedure. Suppose researchers only filter through posts with cyberspace-specific meta-themes such as Accessibility and Anonymity. In that case, there is a chance that they will mislabel a Non-Cyberbullying post as a Cyberbullying post simply because Accessibility and/or Anonymity factors are present and vice versa. This finding indicates that any volunteers or experts who perform labelling on a dataset should be briefed beforehand in taking such cautions when evaluating an SM post based on the six different meta themes.

We learned two lessons for the Characterization of Cyberbullying, otherwise known as the process of Feature Engineering or Representation Learning. The first lesson is that contextual/non-textual features complement textual features. They are not replacements. Secondly, we have also established that using a multimodal/combined feature set, in comparison to single feature sets, greatly improved the performance of ML classifiers. However, there should not be a misconception that any contextual/non-textual feature would suffice in its supplementary action to textual features in increasing the performance of said classifiers. Some contextual/non-textual features are more relevant than others; this is where the importance of Feature Selection comes in, to select only the significant contextual/non-textual features and discard irrelevant ones. Feature Selection has been shown repeatedly in several studies to benefit the forthcoming Classification step significantly by ensuring that the classifiers are trained with the cream of the crop, regardless of whether it is contextual/non-textual or textual features. All in all, researchers looking to: i) perform Feature Engineering with different pre-existing contextual/non-textual features and build new relevant contextual/non-textual features; ii) improve present textual features; iii) use Feature Selection to select the best features by comparing different feature combinations to see which one works the best; and iv) use contextual/non-textual and textual features in a multimodal/combined feature set, never in isolation.

Concerning the Classification of Cyberbullying, the two most important findings are the context-dependent effect on selecting both ML classifiers and the subsequent Evaluation Metrics. Traditional ML approaches appear to outshine the more complex DNN/DL models for modestly-sized datasets. Their training requires large amounts of data, the absence of which will result in overfitting and unrealistically high performance, which is undesirable. This shows that in considering which ML classifiers to deploy in one's system, one should always decide based on the context. As observed in the case of datasets that are not large, complex solutions are not necessarily better. Sometimes it is worth looking at more straightforward approaches before contemplating using more complex ones.

In summary, when performing Classification, selecting ML classifiers and performance metrics best suited for the context is recommended instead of going for what is popular or commonly used. Keeping an open mind and experimenting with different ML classifiers and performance metrics will allow researchers to determine the best classifiers for their specific data context and have metrics that can represent how well their classifiers are performing in that particular context. Under the proper context, DNN/DL approaches outperform Traditional ML classifiers. Therefore, researchers should explore other DNN/DL approaches.

6 Conclusion

This study has demonstrated a deep dive into the state of the art of papers on Cyberbullying detection in social media using ML classifiers through a systematic literature review to provide insights into the three pillars: Conceptualization, Characterization, and Classification of Cyberbullying. Our results show that: 1) Cyberbullying is distinct from Traditional Bullying though they share some commonalities. 2) The combined usage of contextual and textual features in a multimodal feature set can improve Classification accuracy. 3) The selection of ML classifiers and performance metrics is context-dependent. In isolation, these results have been presented in prior works multiple times, so they are not unique to this study. However, this study offers a condensed view of all the results in a continuous narrative from one pillar to another; when observed in separation, the whys and hows may not be adequate for a holistic understanding of the subject matter.

A. Limitations

Suppose there should be one constraint in this study. In that case, it is the fact that the inclusion and exclusion criteria set in the SLR are far too limiting to retrieve the breadth of studies that are essential for making up state of the art on papers related to Cyberbullying detection using ML in its complete totality. For instance, we only considered papers either using Traditional ML classifiers, DNN/DL approaches, or a mix of the two. Furthermore, we excluded those papers that used other soft computing techniques such as Fuzzy Systems, Evolutionary Computing, and Swarm Intelligence. This exclusion potentially hides critical takes on feature engineering and evaluation metrics used in those studies, creating a distorted and narrow worldview for readers.

B. Suggestions for Future Study

Hence, future iterations of this study should take an all-inclusive approach, generalizing the inclusion criteria to allow for more studies to be screened through the SLR. Additionally, we believe that another direction for future studies is to consider using the term Cyberaggression instead of Cyberbullying as an inclusion criterion in the SLR and the title piece of research. Potentially, Cyberaggression might be a more suitable and all-encompassing term than Cyberbullying. It is because it suggests a digital equivalent of Traditional Bullying, which in turn means that they should share similar core criteria, which in the case of Cyberbullying is not so clear cut at times.

Acknowledgement

The authors hereby acknowledge the review support offered by the IJPCC reviewers who took their time to study the manuscript and find it acceptable for publishing.

Conflict of Interest

The authors declare that there is no conflict of Interest.

References

[1] I. Vojinovic. (2021). *Heart-Breaking Cyberbullying Statistics for 2021*. Available: https://dataprot.net/statistics/cyberbullying-statistics/

[2] K. Hellfeldt, L. López-Romero, and H. Andershed, "Cyberbullying and psychological well-being in young adolescence: the potential protective mediation effects of social support from family, friends, and teachers," *International journal of environmental research and public health,* vol. 17, no. 1, p. 45, 2020.

[3] C. L. Nixon, "Current perspectives: the impact of cyberbullying on adolescent health," *Adolescent health, medicine and therapeutics,* vol. 5, p. 143, 2014.

[4] Q. Faryadi, "Cyber bullying and academic performance," *Online Submission,* vol. 1, no. 1, pp. 23-30, 2011.

[5] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *Journal of child psychology and psychiatry,* vol. 49, no. 4, pp. 376-385, 2008.

[6] S. Salawu, Y. He, and J. Lumsden, "BullStop: A Mobile App for Cyberbullying Prevention," in *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, 2020, pp. 70-74.

[7] A. Kumar and N. Sachdeva, "Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis," *Multimedia Tools and Applications,* vol. 78, no. 17, pp. 23973-24010, 2019.

[8] M. Dadvar, F. d. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, 2012: University of Ghent.

[9] S. Berne, A. Frisén, and J. Kling, "Appearance-related cyberbullying: A qualitative investigation of characteristics, content, reasons, and effects," *Body image,* vol. 11, no. 4, pp. 527-533, 2014.

[10] K. Naruskov, P. Luik, A. Nocentini, and E. Menesini, "Estonian Students' Perception and Definition of Cyberbullying," *Trames: A Journal of the Humanities and Social Sciences,* vol. 16, no. 4, p. 323, 2012.

[11] A. Nocentini, J. Calmaestra, A. Schultze-Krumbholz, H. Scheithauer, R. Ortega, and E. Menesini, "Cyberbullying: Labels, behaviours and definition in three European countries," *Journal of Psychologists and Counsellors in Schools,* vol. 20, no. 2, pp. 129-142, 2010.

[12] Ç. Topcu, A. Yıldırım, and Ö. Erdur-Baker, "Cyber bullying@ schools: What do Turkish adolescents think?," *International Journal for the Advancement of Counselling,* vol. 35, no. 2, pp. 139-151, 2013.

[13] H. Vandebosch and K. Van Cleemput, "Defining cyberbullying: A qualitative research into the perceptions of youngsters," *CyberPsychology & Behavior,* vol. 11, no. 4, pp. 499-503, 2008.

[14] L. R. Betts and K. A. Spenser, "'People think it'sa harmless joke": young people's understanding of the impact of technology, digital vulnerability and cyberbullying in the United Kingdom," *Journal of Children and Media,* vol. 11, no. 1, pp. 20-35, 2017.

[15] N. C. Jacobs, L. Goossens, F. Dehue, T. Völlink, and L. Lechner, "Dutch cyberbullying victims' experiences, perceptions, attitudes and motivations related to (coping with) cyberbullying: Focus group interviews," *Societies,* vol. 5, no. 1, pp. 43-64, 2015.

[16] R. Rafferty and T. Vander Ven, "'I hate everything about you": A qualitative examination of cyberbullying and on-line aggression in a college sample," *Deviant behavior,* vol. 35, no. 5, pp. 364-377, 2014.

[17] C. Langos, "Cyberbullying: The challenge to define," *Cyberpsychology, behavior, and social networking,* vol. 15, no. 6, pp. 285-289, 2012.

[18] R. A. Baron, *Human aggression*. New York: Plenum Press, 1977.

[19] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?," *Scandinavian journal of psychology,* vol. 49, no. 2, pp. 147-154, 2008.

[20] F. Mishna, M. Saini, and S. Solomon, "Ongoing and online: Children and youth's perceptions of cyber bullying," *Children and Youth Services Review,* vol. 31, no. 12, pp. 1222-1228, 2009.

[21] W. V. Pelfrey Jr and N. Weber, "Talking smack and the telephone game: Conceptualizing cyberbullying with middle and high school youth," *Journal of Youth Studies,* vol. 17, no. 3, pp. 397-414, 2014.

[22] N. Baas, M. D. De Jong, and C. H. Drossaert, "Children's perspectives on cyberbullying: insights based on participatory research," *Cyberpsychology, behavior, and social networking,* vol. 16, no. 4, pp. 248-253, 2013.

[23] H. S. Abu Bakar, "The emergence themes of cyberbullying among adolescences," *International Journal of Adolescence and Youth,* vol. 20, no. 4, pp. 393-406, 2015.

[24] J. Burnham and V. Wright, "Cyberbullying: What Middle School Students Want You to Know," *Alabama Counseling Association Journal,* vol. 38, no. 1, pp. 3-12, 2012.

[25] S. Hinduja and J. W. Patchin, *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin press, 2014.

[26] C. Katzer, D. Fetchenhauer, and F. Belschak, "Cyberbullying: Who are the victims?: A comparison of victimization in internet chatrooms and victimization in school," *Journal of Media Psychology: Theories, Methods, and Applications,* vol. 21, no. 1, p. 25, 2009.

[27] J. J. Dooley, J. Pyżalski, and D. Cross, "Cyberbullying versus face-to-face bullying: A theoretical and conceptual review," *Zeitschrift für Psychologie/Journal of Psychology,* vol. 217, no. 4, p. 182, 2009.

[28] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing,* vol. 11, no. 1, pp. 3-24, 2017.

[29] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2011, vol. 5, no. 3, pp. 11-17.

[30] P. J. C. Pérez, C. J. L. Valdez, M. d. G. C. Ortiz, J. P. S. Barrera, and P. F. Pérez, "MISAAC: Instant messaging tool for ciberbullying detection," in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, 2012, p. 1: The Steering Committee of The World Congress in Computer Science, Computer … .

[31] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," *Communications in information science and management engineering,* vol. 3, no. 5, p. 238, 2013.

[32] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in Vine," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2015, pp. 617-622: IEEE.
[33] M. Munezero, C. S. Montero, T. Kakkonen, E. Sutinen, M. Mozgovoy, and V. Klyuev, "Automatic detection of antisocial behaviour in texts," *Informatica,* vol. 38, no. 1, 2014.
[34] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," *Proceedings of the Content Analysis in the WEB,* vol. 2, pp. 1-7, 2009.
[35] S. O. Sood, J. Antin, and E. Churchill, "Using crowdsourcing to improve profanity detection," in *2012 AAAI Spring Symposium Series*, 2012.
[36] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, 2014, pp. 3-6.
[37] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in *International semantic web conference*, 2012, pp. 508-524: Springer.
[38] N. Oliveira, P. Cortez, and N. Areal, "On the predictability of stock market behavior using stocktwits sentiment and posting volume," in *Portuguese conference on artificial intelligence*, 2013, pp. 355-365: Springer.
[39] H. Sanchez and S. Kumar, "Twitter bullying detection," *ser. NSDI,* vol. 12, no. 2011, p. 15, 2011.
[40] J. Sheeba and K. Vivekanandan, "Low frequency keyword extraction with sentiment classification and cyberbully detection using fuzzy logic technique," in *2013 IEEE International Conference on Computational Intelligence and Computing Research*, 2013, pp. 1-5: IEEE.
[41] L. P. Del Bosque and S. E. Garza, "Aggressive text detection for cyberbullying," in *Mexican International Conference on Artificial Intelligence*, 2014, pp. 221-232: Springer.
[42] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment analysis for effective detection of cyber bullying," in *Asia-Pacific Web Conference*, 2012, pp. 767-774: Springer.
[43] J.-M. Xu, X. Zhu, and A. Bellmore, "Fast learning for sentiment analysis on bullying," in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2012, pp. 1-6.
[44] M. Munezero, M. Mozgovoy, T. Kakkonen, V. Klyuev, and E. Sutinen, "Antisocial behavior corpus for harmful language detection," in *2013 Federated Conference on Computer Science and Information Systems*, 2013, pp. 261-265: IEEE.
[45] S. M. Serra and H. S. Venter, "Mobile cyber-bullying: A proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness," in *2011 Information Security for South Africa*, 2011, pp. 1-5: IEEE.
[46] M. Dadvar and F. De Jong, "Cyberbullying detection: a step toward a safer internet yard," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 121-126.
[47] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," in *Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015*, 2015, pp. 280-285.
[48] G. NaliniPriya and M. Asswini, "A dynamic cognitive system for automatic detection and prevention of cyber-bullying attacks," *ARPN J. Eng. Appl. Sci,* vol. 10, no. 10, pp. 4618-4626, 2015.
[49] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. S. Turaga, "Learning Feature Engineering for Classification," in *Ijcai*, 2017, pp. 2529-2535.
[50] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 2012, pp. 71-80: IEEE.
[51] B. S. Nandhini and J. Sheeba, "Online social network bullying detection using intelligence techniques," *Procedia Computer Science,* vol. 45, pp. 485-492, 2015.
[52] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Transactions on Affective Computing,* vol. 8, no. 3, pp. 328-339, 2016.
[53] H. Hosseinmardi, S. A. Mattson, R. Rafiq, R. Han, Q. Lv, and S. Mishra, "Poster: Detection of cyberbullying in a mobile social network: Systems issues," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, 2015, pp. 481-481.
[54] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," in *2015 IEEE international conference on electro/information technology (EIT)*, 2015, pp. 611-616: IEEE.
[55] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying," *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13,* vol. 239, pp. 319-428, 2014.
[56] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," in *2016 23rd International conference on pattern recognition (ICPR)*, 2016, pp. 432-437: IEEE.
[57] L. Cheng, K. Shu, S. Wu, Y. N. Silva, D. L. Hall, and H. Liu, "Unsupervised cyberbullying detection via time-informed gaussian mixture model," *arXiv preprint arXiv:2008.02642,* 2020.
[58] B. Pranto, S. M. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, and S. Momen, "Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh," *Information,* vol. 11, no. 8, p. 374, 2020.
[59] M. Suresh and R. Anitha, "Evaluating machine learning algorithms for detecting DDoS attacks," in *International Conference on Network Security and Applications*, 2011, pp. 441-452: Springer.
[60] J. P. Higgins and S. Green, "Cochrane handbook for systematic reviews of interventions. Chichester, England; Hoboken," ed: NJ: Wiley-Blackwell, 2008.
[61] W. Mengist, T. Soromessa, and G. Legese, "Method for conducting systematic literature review and meta-analysis for environmental science research," *MethodsX,* vol. 7, p. 100777, 2020.
[62] D. Moher *et al.*, "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement," *Systematic reviews,* vol. 4, no. 1, pp. 1-9, 2015.
[63] I. F. del Amo, J. A. Erkoyuncu, R. Roy, R. Palmarini, and D. Onoufriou, "A systematic review of Augmented Reality content-related techniques for knowledge transfer in maintenance applications," *Computers in Industry,* vol. 103, pp. 47-71, 2018.
[64] S. Sepúlveda, M. Diéguez, G. Farías, and C. Cachero, "Systematic literature review protocol. Learning-outcomes and teaching-learning process: a Bloom's taxonomy perspective," *arXiv preprint arXiv:1911.09489,* 2019.
[65] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.
[66] M. Unterkalmsteiner, T. Gorschek, A. M. Islam, C. K. Cheng, R. B. Permadi, and R. Feldt, "Evaluation and measurement of software process improvement—a systematic literature review," *IEEE Transactions on Software Engineering,* vol. 38, no. 2, pp. 398-424, 2011.
[67] L. Chen and M. A. Babar, "A systematic review of evaluation of variability management approaches in software product lines," *Information and Software Technology,* vol. 53, no. 4, pp. 344-362, 2011.
[68] S. A. Hemphill *et al.*, "Longitudinal predictors of cyber and traditional bullying perpetration in Australian secondary school students," *Journal of Adolescent Health,* vol. 51, no. 1, pp. 59-65, 2012.
[69] A. Ševčíková, D. Šmahel, and M. Otavová, "The perception of cyberbullying in adolescent victims," *Emotional and behavioural difficulties,* vol. 17, no. 3-4, pp. 319-328, 2012.
[70] W. Cassidy, C. Faucher, and M. Jackson, "Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice," *School psychology international,* vol. 34, no. 6, pp. 575-612, 2013.

[71] S. Eden, T. Heiman, and D. Olenik-Shemesh, "Teachers' perceptions, beliefs and concerns about cyberbullying," *British journal of educational technology,* vol. 44, no. 6, pp. 1036-1052, 2013.
[72] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth," *Psychological bulletin,* vol. 140, no. 4, p. 1073, 2014.
[73] H. J. Thomas, J. P. Connor, and J. G. Scott, "Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents–a review," *Educational psychology review,* vol. 27, no. 1, pp. 135-152, 2014.
[74] I. Cuadrado-Gordillo and I. Fernández-Antelo, "Adolescents' perception of the characterizing dimensions of cyberbullying: Differentiation between bullies' and victims' perceptions," *Computers in Human Behavior,* vol. 55, pp. 653-663, 2016.
[75] A. E. Fahy, S. A. Stansfeld, M. Smuk, N. R. Smith, S. Cummins, and C. Clark, "Longitudinal associations between cyberbullying involvement and adolescent mental health," *Journal of Adolescent Health,* vol. 59, no. 5, pp. 502-509, 2016.
[76] R. Dennehy, S. Meaney, K. A. Walsh, C. Sinnott, M. Cronin, and E. Arensman, "Young people's conceptualizations of the nature of cyberbullying: A systematic review and synthesis of qualitative research," *Aggression and violent behavior,* vol. 51, p. 101379, 2020.
[77] O. Aluede, F. Adeleke, D. Omoike, and J. Afen-Akpaida, "A review of the extent, nature, characteristics and effects of bullying behaviour in schools," *Journal of Instructional Psychology,* vol. 35, no. 2, p. 151, 2008.
[78] K. L. Modecki, J. Minchin, A. G. Harbaugh, N. G. Guerra, and K. C. Runions, "Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying," *Journal of Adolescent Health,* vol. 55, no. 5, pp. 602-611, 2014.
[79] I. Tsaousis, "The relationship of self-esteem to bullying perpetration and peer victimization among schoolchildren and adolescents: A meta-analytic review," *Aggression and violent behavior,* vol. 31, pp. 186-199, 2016.
[80] D. Salin *et al.*, "Workplace bullying across the globe: A cross-cultural comparison," *Personnel Review,* 2018.
[81] J. Stuart and N. Szeszeran, "Bullying in the military: a review of the research on predictors and outcomes of bullying victimization and perpetration," *Military Behavioral Health,* vol. 9, no. 3, pp. 255-266, 2020.
[82] N. Rezvani, A. Beheshti, and A. Tabebordbar, "Linking textual and contextual features for intelligent cyberbullying detection in social media," in *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*, 2020, pp. 3-10.
[83] J. Bayzick, A. Kontostathis, and L. Edwards, "Detecting the presence of cyberbullying using computer software," 2011.
[84] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine learning and applications and workshops*, 2011, vol. 2, pp. 241-244: IEEE.
[85] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS),* vol. 2, no. 3, pp. 1-30, 2012.
[86] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context," in *European conference on information retrieval*, 2013, pp. 693-696: Springer.
[87] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: query terms and techniques," in *Proceedings of the 5th annual acm web science conference*, 2013, pp. 195-204.
[88] V. Nahar, S. Al-Maskari, X. Li, and C. Pang, "Semi-supervised learning for cyberbullying detection in social networks," in *Australasian Database Conference*, 2014, pp. 160-171: Springer.
[89] M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, and K. Araki, "Extracting patterns of harmful expressions for cyberbullying detection," in *Proceedings of 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'15), The First Workshop on Processing Emotions, Decisions and Opinions*, 2015, pp. 370-375.
[90] C. Van Hee *et al.*, "Automatic detection and prevention of cyberbullying," in *International Conference on Human and Social Analytics (HUSO 2015)*, 2015a, pp. 13-18: IARIA.
[91] C. Van Hee *et al.*, "Detection and fine-grained classification of cyberbullying events," in *Proceedings of the international conference recent advances in natural language processing*, 2015b, pp. 672-680.
[92] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Computers in Human Behavior,* vol. 63, pp. 433-443, 2016.
[93] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th international conference on distributed computing and networking*, 2016, pp. 1-6.
[94] H. J. Escalante, E. Villatoro-Tello, S. E. Garza, A. P. López-Monroy, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "Early detection of deception and aggressiveness using profile-based representations," *Expert Systems with Applications,* vol. 89, pp. 99-111, 2017.
[95] B. Haidar, M. Chamoun, and A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," *Advances in Science, Technology and Engineering Systems Journal,* vol. 2, no. 6, pp. 275-284, 2017.
[96] S. A. Özel, E. Saraç, S. Akdemir, and H. Aksu, "Detection of cyberbullying on social media messages in Turkish," in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 366-370: IEEE.
[97] E. Raisi and B. Huang, "Cyberbullying detection with weakly supervised machine learning," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 409-416.
[98] V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 2090-2099.
[99] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur, "Using fuzzy fingerprints for cyberbullying detection in social networks," in *2018 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, 2018, pp. 1-7: IEEE.
[100] J. Hani, N. Mohamed, M. Ahmed, Z. Emad, E. Amer, and M. Ammar, "Social media cyberbullying detection using machine learning," *International Journal of Advanced Computer Science and Applications,* vol. 10, no. 5, 2019.
[101] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using Twitter users' psychological features and machine learning," *Computers & Security,* vol. 90, p. 101710, 2020.
[102] O. Gencoglu, "Cyberbullying detection with fairness constraints," *IEEE Internet Computing,* vol. 25, no. 1, pp. 20-29, 2020.
[103] S. M. Kargutkar and V. Chitre, "A study of cyberbullying detection using machine learning techniques," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 734-739: IEEE.
[104] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on Twitter," *Future Internet,* vol. 12, no. 11, p. 187, 2020.
[105] E. C. Ates, E. Bostanci, and M. S. Guzel, "Comparative performance of machine learning algorithms in cyberbullying detection: Using turkish language preprocessing techniques," *arXiv preprint arXiv:2101.12718,* 2021.
[106] A. Bozyiğit, S. Utku, and E. Nasibov, "Cyberbullying detection: Utilizing social media features," *Expert Systems with Applications,* vol. 179, p. 115001, 2021.

[107] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Computer Science,* vol. 181, pp. 605-611, 2021.
[108] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford,* vol. 1, no. 12, p. 2009, 2009.
[109] V. K. Singh, Q. Huang, and P. K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 884-887: IEEE.
[110] M. Sintaha, S. B. Satter, N. Zawad, C. Swarnaker, and A. Hassan, "Cyberbullying detection using sentiment analysis in social media," BRAC University, 2016.
[111] W. Romsaiyud, K. na Nakornphanom, P. Prasertsilp, P. Nurarak, and P. Konglerd, "Automated cyberbullying detection using clustering appearance patterns," in *2017 9th International Conference on Knowledge and smart Technology (KST)*, 2017, pp. 242-247: IEEE.
[112] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European conference on information retrieval*, 2018, pp. 141-153: Springer.
[113] H. Nurrahmi and D. Nurjanah, "Indonesian twitter cyberbullying detection using text classification and user credibility," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 543-548: IEEE.
[114] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," *Proceedings of the ACM on Human-Computer Interaction,* vol. 2, no. CSCW, pp. 1-26, 2018.
[115] N. Tahmasbi and E. Rastegari, "A socio-contextual approach in automated detection of cyberbullying," 2018.
[116] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of cyberbullying using deep neural network," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019, pp. 604-607: IEEE.
[117] M. Biesek, "Comparison of Traditional Machine Learning Approach and Deep Learning Models in Automatic Cyberbullying Detection for Polish Language," in *Proceedings of the PolEval 2019 Workshop*, 2019, pp. 121-126.
[118] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 339-347.
[119] J. A. Cornel *et al.*, "Cyberbullying detection for online games chat logs using deep learning," in *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 2019, pp. 1-5: IEEE.
[120] B. Haidar, M. Chamoun, and A. Serhrouchni, "Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning," in *2019 international conference on internet of things (ithings) and ieee green computing and communications (greencom) and ieee cyber, physical and social computing (cpscom) and ieee smart data (smartdata)*, 2019, pp. 323-327: IEEE.
[121] A. Kumar, S. Nayak, and N. Chandra, "Empirical analysis of supervised machine learning techniques for cyberbullying detection," in *International Conference on Innovative Computing and Communications*, 2019, pp. 223-230: Springer.
[122] Y. Liu, P. Zavarsky, and Y. Malik, "Non-linguistic features for cyberbullying detection on a social media platform using machine learning," in *International Symposium on Cyberspace Safety and Security*, 2019, pp. 391-406: Springer.
[123] M. Yao, C. Chelmis, and D.-S. Zois, "Cyberbullying ends here: Towards robust detection of cyberbullying in social media," in *The World Wide Web Conference*, 2019, pp. 3427-3433.
[124] J. Zhang, T. Otomo, L. Li, and S. Nakajima, "Cyberbullying detection on twitter using multiple textual features," in *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, 2019, pp. 1-6: IEEE.
[125] J. Alasadi, R. Arunachalam, P. K. Atrey, and V. K. Singh, "A fairness-aware fusion framework for multimodal cyberbullying detection," in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 166-173: IEEE.
[126] A. Ali and A. M. Syed, "Cyberbullying detection using machine learning," *Pakistan Journal of Engineering and Technology,* vol. 3, no. 2, pp. 45-50, 2020.
[127] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, "Cyberbullying detection on social networks using machine learning approaches," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1-6: IEEE.
[128] A. Kumar and N. Sachdeva, "Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data," *Multimedia systems,* pp. 1-15, 2020.
[129] D. Van Bruwaene, Q. Huang, and D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media," *Language Resources and Evaluation,* vol. 54, no. 4, pp. 851-874, 2020.
[130] K. Wang, Q. Xiong, C. Wu, M. Gao, and Y. Yu, "Multi-modal cyberbullying detection on social networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8: IEEE.
[131] M. F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. N. Ryen, A. Hossain, and F. B. Ashraf, "Cyberbullying detection using deep neural network from social media comments in bangla language," *arXiv preprint arXiv:2106.04506,* 2021.
[132] T. Alsubait and D. Alfageh, "Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments," *International Journal of Computer Science & Network Security,* vol. 21, no. 1, pp. 1-5, 2021.
[133] N. A. Azeez, S. O. Idiakose, C. J. Onyema, and C. Van Der Vyver, "Cyberbullying Detection in Social Networks: Artificial Intelligence Approach," *Journal of Cyber Security and Mobility,* pp. 745–774-745–774, 2021.
[134] J. Eronen, M. Ptaszynski, F. Masui, A. Smywiński-Pohl, G. Leliwa, and M. Wroczynski, "Improving classifier training efficiency for automatic cyberbullying detection with feature density," *Information Processing & Management,* vol. 58, no. 5, p. 102616, 2021.
[135] S. Ge, L. Cheng, and H. Liu, "Improving cyberbullying detection with user interaction," in *Proceedings of the Web Conference 2021*, 2021, pp. 496-506.
[136] R. Ghosh, S. Nowal, and G. Manju, "Social media cyberbullying detection using machine learning in bengali language," *Int J Eng Res Technol,* 2021.
[137] W. J. Chong, H. N. Chua and M. F. Gan, "Comparing Zero-Shot Text Classification and Rule-Based Matching in Identifying Cyberbullying Behaviors on Social Media," 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 2022, pp. 1-5, doi: 10.1109/IICAIET55139.2022.9936821.