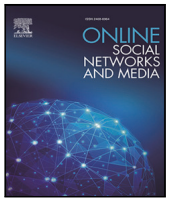
Online Social Networks and Media 36 (2023) 100250

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem



Session-based cyberbullying detection in social media: A survey

Peiling Yi *, Arkaitz Zubiaga

School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London, E1 4NS, United Kingdom

ARTICLE INFO

Keywords:
Cyberbullying detection
Cyberbullying
Session-based cyberbullying detection
Social media
Machine learning
Natural language processing

ABSTRACT

Cyberbullying is a pervasive problem in online social media, where a bully abuses a victim through a social media session. By investigating cyberbullying perpetrated through social media sessions, recent research has looked into mining patterns and features for modelling and understanding the two defining characteristics of cyberbullying: repetitive behaviour and power imbalance. In this survey paper, we define a framework that encapsulates four different steps session-based cyberbullying detection should go through, and discuss the multiple challenges that differ from single text-based cyberbullying detection. Based on this framework, we provide a comprehensive overview of session-based cyberbullying detection in social media, delving into existing efforts from a data and methodological perspective. Our review leads us to proposing evidence-based criteria for a set of best practices to create session-based cyberbullying datasets. In addition, we perform benchmark experiments comparing the performance of state-of-the-art session-based cyberbullying detection models as well as large pre-trained language models across two different datasets. Through our review, we also put forth a set of open challenges as future research directions.

1. Introduction

"Bullying" is defined as the repeated and deliberate aggressive behaviour by a group or individual towards a person who is in a more vulnerable position to defend themselves [1]. Cyberbullying is widely defined as a form of bullying that is perpetrated through online devices [2], which may be an individual or group sending, posting, or sharing negative, harmful, false, or mean content about someone else. Some cyberbullying crosses the line of unlawful or criminal behaviour [3,4]. The concern caused by the increasing number of teen suicides linked to cyberbullying incidents led lawmakers and politicians to consider new criminal legislation adapted to cyberbullying. However, research into how police and law enforcement officials can better respond to cyberbullying incidents is still limited. While some cyberbullying incidents warrant criminal charges in certain circumstances, these incidents go beyond "bullying" and fall under other legal categories [5–7]. Thus, the definition of cyberbullying used in this study relies on widely accepted definitions from a range of relevant research fields.

There are two inherent characteristics of cyberbullying that are consistently referred to: repeated aggression and power imbalance, both of which are key aspects to identifying cases of cyberbullying behaviour [8–11]. Repeated aggression means that it happens more than just once or twice. Thus, a one-off cyberbullying-like message is not deemed cyberbullying [12–14]. The "imbalance of power" criterion refers to the situation where victims cannot easily defend themselves. Examples could include one user being more technical than another [15], a group of users targeting a user, a popular user targeting a less popular user [12], or the anonymity of a bully [16]. In this study, we delve into the investigation of how these two key factors are being considered in the current literature.

Cyberbullying detection is the task of automatically identifying cyberbullying events from online data, with the aim of stopping the abuse and preventing further harm [17,18]. Developing an ability to detect cyberbullying events is however challenging, as it needs to capture the recurrent nature of the abusive behaviour. An ability to detect offensive or toxic sentences, as in for example hate speech detection and abusive language detection [19], does not suffice; ideally, it needs to consider the full history of the conversation (i.e. a social media session) to identify the recurrent nature inherent to cyberbullying events [20], by constructing a representation of the interaction between the bully and the victim.

Fig. 1 illustrates an example of an act of cyberbullying in an online chat [21,22]. A bully recurrently sends mocking messages that can sometimes reveal personal or sensitive information of an indefensible victim. Messages by both the victim and the bully may contain offensive words, where however the victim will generally be using such words to try to defend themselves from the bully. Thus modelling the session as a whole or as isolated posts can indeed make a difference. A single-post cyberbullying detection model may flag a defensive sentence as a case

* Corresponding author.
E-mail addresses: p.yi@qmul.ac.uk (P. Yi), a.zubiaga@qmul.ac.uk (A. Zubiaga).
URL: http://www.zubiaga.org/ (A. Zubiaga).

https://doi.org/10.1016/j.osnem.2023.100250
Received 8 August 2022; Received in revised form 22 May 2023; Accepted 24 May 2023
Available online 17 June 2023

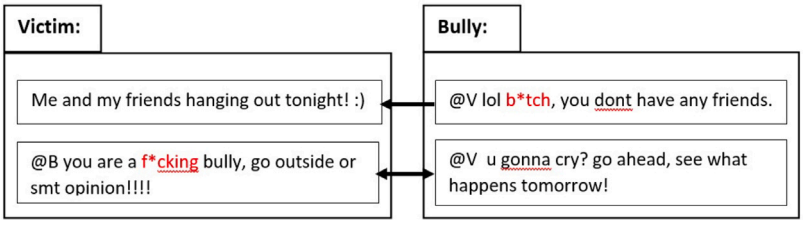


Fig. 1. Example of a case of cyberbullying.

of cyberbullying due to its offensive words. However, in cases such as "u gonna cry? go ahead, see what happens tomorrow", there are no offensive words; however, it could contain a case of cyberbullying if the surrounding social media session indicates so.

The task of session-based cyberbullying detection consists in determining if cyberbullying incidents are present in a social media session. A social media session typically consists of an initial post/image/video, and a series of related comments involving user interaction, user information, spatial location, and other social content [23]. There are important aspects that a social media session can provide which cannot be inferred from isolated posts, as the aggregation of the broader context is needed. Performing cyberbullying detection by modelling social media sessions provides a holistic view of the power imbalance between the bully and the victim, which cannot be inferred from the limited information available in isolated posts. In addition, the repetitive nature of cyberbullying can only be captured by the sequence of comments in a conversation session.

In this survey paper, we delve into the current development of research into cyberbullying detection in social media, with a particular focus on methods incorporating social media sessions into their pipeline. A number of recent surveys have covered cyberbullying detection, which have however had different foci from the one here on social media sessions. Existing surveys have predominantly covered cyberbullying detection in general without a focus on social media sessions [24–28], whereas others have focused on more specific aspects including a critical review on the definitions and operationalisation of cyberbullying [29], an overview of the implications of cyberbullying [24] and providing a taxonomy of the different types of cyber-attacks [30].

We survey 10 publicly accessible cyberbullying datasets and 55 cyberbullying detection models by examining how they adhere to the criteria set out above, which we formalise in a cyberbullying detection framework. We refer to this framework as the Social Media Session-Based Cyberbullying Detection (SSCD) that unifies the definition, data collection, and detection of cyberbullying.

Our survey paper makes a number of contributions, of which we highlight:

- We define the four steps of SSCD: (i) social media platform selection, (ii) session-based data collection, (iii) cyberbullying annotation, and (iv) session-based cyberbullying detection.
- We give an overview of the existing datasets and methods in accordance with the SSCD framework.
- We define a set of evidence-based criteria recommended for the selection and creation of an SSCD dataset.
- We perform experiments investigating the use of two state-of-the-art session-based cyberbullying detection models and nine different pre-trained language models to tackle SSCD tasks.
- Informed by our literature review on existing datasets and methods, we provide a set of suggestions for consideration in future work for dataset creation, model development as well as reporting in scientific publications.
- We provide a comprehensive understanding of cyberbullying detection at the social media session level from a data and methodological perspective.

Paper structure.. The aim of the survey is to provide a comprehensive understanding of cyberbullying detection at the social media session level from a data and methodological perspective. In the next section, we describe the methodology we follow to conduct this survey as well as how the relevant research papers were selected. Then, in Section 3, we introduce and describe the SSCD framework that defines the modelling of the cyberbullying detection task considering social media sessions. Section 4 discusses existing datasets, for which we analyse 10 cyberbullying detection datasets, discussing their data collection and annotation strategies. Then, we provide a set of recommendations for creating SSCD-based Datasets in Section 5. In Sections 6 and 7 we then discuss existing methods and research directions in cyberbullying detection as well as existing efforts on modelling the task in line with the SSCD framework. In Section 8, we continue by presenting experiments with state-of-the-art language models. After that, we discuss open challenges within session-based cyberbullying detection and conclude the paper in Section 9.

2. Survey methodology

In this section, we describe in detail how we adopt the PRISMA framework [31] to filter significant relevant studies. Then, we combine the framework of the data statement [32] and the guidelines for writing systematic reviews [33] to reduce research bias.

2.1. Selection of studies

We followed a consistent methodology to retrieve relevant papers to be covered in our survey, with the aim of ensuring good coverage of papers while also avoiding biased selection based on subjective criteria. We used four different keywords (i.e. 'cyberbullying detection', 'bullying detection', 'cyberbullying datasets', and 'cyberbullying software') to retrieve publications from a wide set of scientific search engines, including Google Scholar, ACM, IEEE, and arXiv. We then reviewed the included studies according to the 2020 PRISMA Update Study Flowchart [31]. The PRISMA flowchart visually shows the changes in the number of articles at different stages, making the selection process transparent by reporting decisions made throughout the review process. The process is shown in Fig. 2.

After retrieving all these papers and removing duplicates, 217 publications were selected for more careful analysis and validation. A final set of 55 publications was selected after removing those that did not fall into any of the following exclusion criteria:

- *Exclusion criteria 1:* Cyberbullying detection is not applied on social media platforms, hence it is outside of our scope.
- *Exclusion criteria 2:* While the study is about or mentions cyberbullying detection, there is no implementation, study, or evaluation of a detection model, e.g. papers discussing how cyberbullying detection can inform policy.
- *Exclusion criteria 3:* The study is a review, survey, or study of theoretical concepts about cyberbullying or cyberbullying detection, without any empirical implementation or analysis.

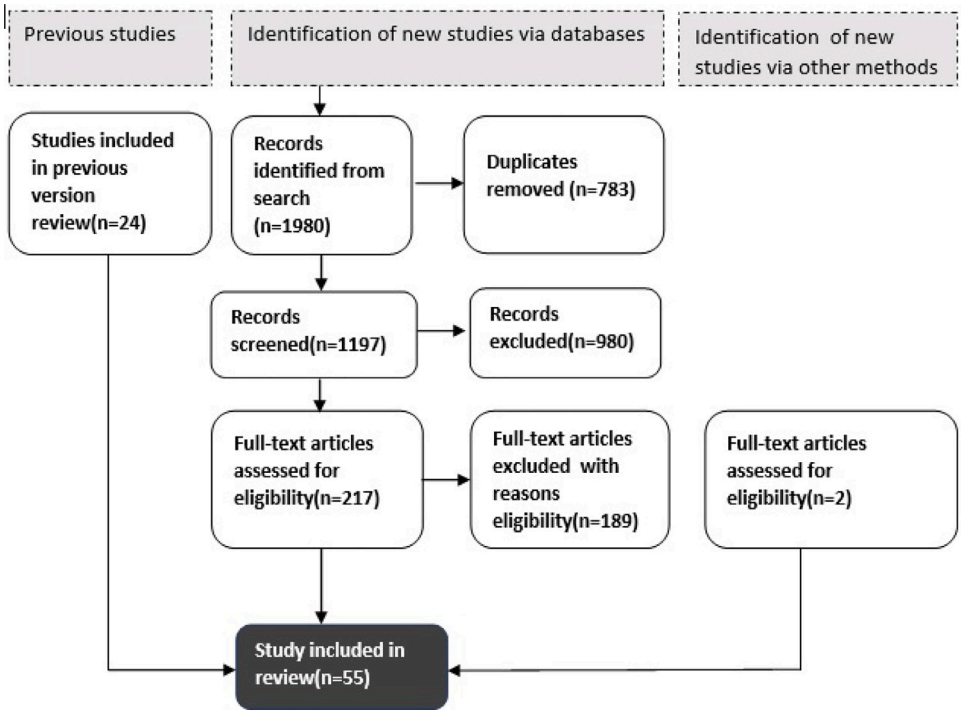


Fig. 2. PRISMA Model for Cyberbullying Detection Research.

Some of the publications matching exclusion criteria 2 or 3 are either discussed throughout this paper or cited for backing up some of our statements, however, they do not conform to the set of studies used for our core discussion of datasets and models for cyberbullying detection.

2.2. Approach for study analysis

The main aim of this survey paper is to provide a comprehensive understanding of SSCD from the perspectives of data and methodologies and combine both to perform a critical analysis of SSCD. To do so, in this study we adopt the "data statements" framework [32]. Data statements suggest how datasets should be created in Natural Language Processing (NLP) research, with the aim of increasing transparency and helping alleviate issues related to exclusion and bias [34]. We also follow guidelines provided by [33] for writing systematic reviews on the subject of software engineering, which we adapt to the field of NLP and cyberbullying detection.

3. SSCD framework

Social media sessions are ubiquitous ecosystems of cyberbullying. In this section, we operationalise this through what we name the Social Media Session-Based Cyberbullying Detection (SSCD) framework that defines the modelling of the cyberbullying detection task considering social media sessions.

Fig. 3 illustrates the structure of the SSCD framework, which consists of four main steps:

1. *Social media platform selection:* The starting point consists in choosing the social media platforms to be considered in the data collection.
2. *Session-based data collection:* It differs from the collection of individual posts in that the unit being collected is an entire social media session, and so is the unit that is annotated for cyberbullying detection. One may also distinguish two types of social media sessions: (i) conversation sessions, which are text-based sessions involving at least two users, as in Fig. 1, and (ii) media sessions, which include other types of media beyond just text, as in Fig. 4.
3. *Cyberbullying annotation:* Where it is crucial to provide detailed and clear criteria for defining what constitutes a case of cyberbullying, paying special attention to the repeated nature and imbalance of power in the incidents.
4. *Session-based cyberbullying detection:* Where the model for cyberbullying detection is built and evaluated.

Fig. 4 illustrates an example of SSCD, based on the collection strategy followed by [35], who constructed a media session-based cyberbullying dataset on Vine, a video-sharing platform. Cyberbullying can happen on Vine in a number of ways, such as posting offensive comments, re-editing, or transcribing someone's video for mockery. The collection and study of video-based social networking sessions were first done by [35]. They defined a social media session in Vine as the posting of a video with its associated likes and comments, restricting collection in this case to a minimum of 15 comments in order for the annotator to have enough context to assess the frequency/repetition of profanity and imbalance power that fits the definition of cyberbullying. Annotators were trained prior to their participation and were given clear instructions explaining the distinctions between cyberaggression and cyberbullying along with a sample of media sessions.

Throughout the paper, we will refer back to the SSCD framework in Fig. 3, linking to the relevant parts.

4. Datasets

In this section, based on the SSCD framework, we discuss 10 cyberbullying datasets from two aspects of data collection and annotation strategies and provide references for whether these datasets meet SSCD criteria.

Among the publications considered within this survey, we selected all those with publicly available or otherwise accessible (e.g. by contacting authors) datasets, the ones that enable further research as well as allow an exploration of the datasets. We gathered a total of 10 datasets that we discuss here. We present key information and statistics for each of these datasets in Table 1. In addition to general information

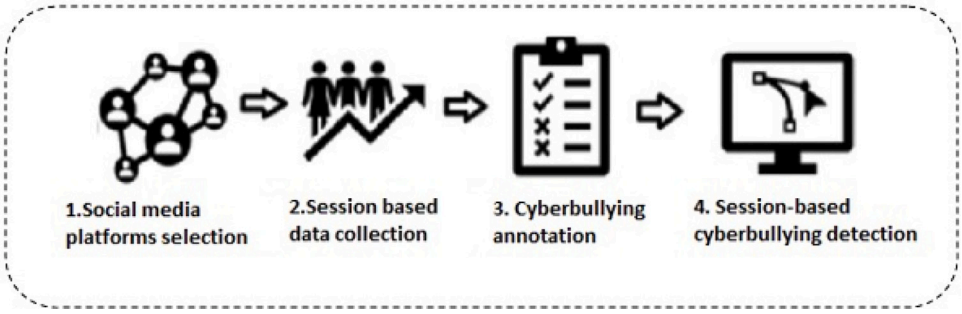


Fig. 3. A general framework for social media session-based cyberbullying detection (SSCD).

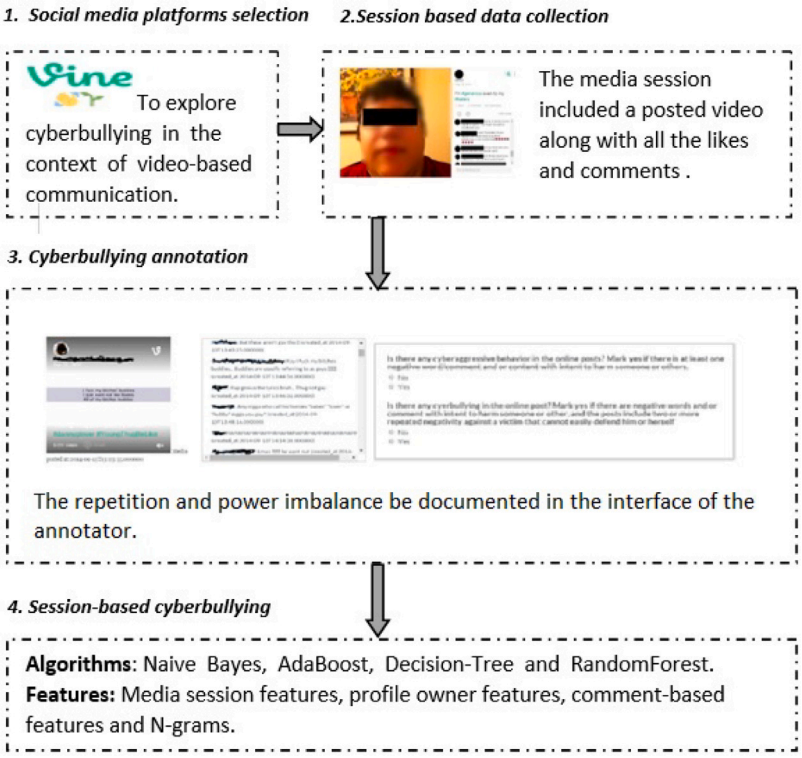


Fig. 4. An instance of SSCD framework. Snapshots are from [35].

such as dataset size, source, and annotation methodology, we also include two additional tables based on SSCD criteria: (i) *Session_based:* Whether the dataset was collected with social media sessions as units, and (ii) *Rigorous Definition:*Whether the data collection and annotation followed a rigorous definition of cyberbullying, with a particular focus on the presence of the two characteristics: repetition and power-imbalance. Datasets adhering to both these criteria are highlighted in bold in the table.

Looking at the statistics and characteristics of these datasets, we make a few observations next:

Datasets are diverse. If we look at the type of data and source platform used to collect the datasets, we see that they are rich in diversity. Across the 10 datasets, as many as 8 different social media platforms were used as sources, where the only platform with more than one dataset is Twitter. The main benefit of this is that it enables further investigation into the problem across different platforms, enabling in turn development of more generalisable models that can detect acts of cyberbullying in different environments. However, there has been little effort to develop more of these datasets in recent years (e.g. 8 datasets were created between 2011–2017, whereas only 2 from 2018–2022).

Datasets are generally imbalanced. According to Table 1, eight of the datasets have a class imbalance where fewer than 30% of the samples belong to the cyberbullying class. This imbalance is a challenge as it has been widely shown to affect the predictive power of machine learning classifiers [45] as shown in experimental reports from previous studies [10,46,47]. It is worth mentioning that some datasets offer the possibility for finer-grained classification such as sexism or racism. However, in the interest of focus and consistency, here we focus on binary classification.

Varying dataset sizes. Dataset sizes vary significantly, from the Vine dataset containing 970 samples to one of the Twitter datasets containing over 534K samples. While datasets generally contain over 10K samples, there is a clear difference in size for the session-based datasets, containing 970, 2.2K, and 4.8K samples, which are understandably smaller given the increased cost of labelling entire sessions.

Limited availability of SSCD datasets. According to our analysis following the SSCD framework, only 6 of the datasets are collected based on sessions, and 4 datasets are labelled following a rigorous definition of cyberbullying. Overall, only 3 of the datasets satisfy both criteria.

Table 1
Available cyberbullying datasets. Datasets collected based on social media sessions and following a rigorous definition of cyberbullying are highlighted in bold.

Dataset statistics							SSCD Datasets Criteria	
Platforms	Size	Ration	Year	Annotation	Collection	Source	Session_based	Rigorous Definition
FormSpring [36]	12,773	0.08	2011	Crowd-sourcing	Crawled	chatcoder.com	Yes	No
Myspace [37]	**4,813**	**0.21**	**2011**	**Research assistant**	**Crawled**	**chatcoder.com**	**Yes**	**Yes**
YouTube[38]	3,468	0.14	2014	Research assistant	Crawled	Figshare.com	No	No
Instagram [39]	**2,218**	**0.29**	**2015**	**Crowd-sourcing**	**Crawled**	**cucybersafety.org**	**Yes**	**Yes**
Vine [40]	**970**	**0.31**	**2015**	**Crowd-sourcing**	**Crawled**	**cucybersafety.org**	**Yes**	**Yes**
Twitter [41]	534,950	0.29	2015	N/A	Twitter API	chatcoder.com	No	No
Wikipedia [42]	115,864	0.11	2017	Crowd-sourcing	Crawled	github.com/sweta20	Yes	No
Twitter [43]	16,090	0.32	2017	Crowd-sourcing	Twitter API	github.com/sweta20	No	Yes
ASKfm [22]	90,296	0.15	2018	Crowd-sourcing	Crawled	cucybersafety.org	Yes	No
Twitter [44]	47,000	0.16	2020	N/A	Twitter API	kaggle.com	No	No

Mismatch between reported and published datasets. The dataset size as stated in the original paper statement does not always match the size of the available dataset. What we report here is the size of the available dataset.

Predominantly crowdsourced annotation of datasets. The majority of the datasets (at least 6 out of 10) used crowdsourcing to annotate the datasets, with only two datasets relying on research assistants with expertise on the subject. Where expertise is important for a difficult annotation task like this, this calls for more datasets using trained annotators.

4.1. Selection of social media platforms

The starting point for constructing a SSCD dataset is the selection of a social media platform. This selection can be motivated by the objectives of the research, which can in turn inform how and what data to collect from the platform of choice, as well as the annotation instructions to provide to annotators.

In existing research, we observe that there have been predominantly two main reasons that motivated the choice of a social media platform:

- *Platforms that are prone to cyberbullying events:* Given the difficulty of retrieving cyberbullying events (i.e. a type of event that can be considered overall rare if we look at all the content in a platform), researchers often turn to platforms that are known to more frequently experience cyberbullying events. The choice of a platform on this basis can be motivated, for example, by the proportion of adolescents and college students known to be users of the platform [36,39,48]. Another feature is the anonymity allowed by certain platforms, which can also indicate higher presence of cyberbullying events in the platform [36]. This is the case of Formspring.me, a Q&A platform where users invite others to ask and answer questions anonymously.
- *Platforms with no existing/public datasets:* Another motivation to choose a particular social media platform has likely been the lack of existing datasets for that particular platform, which has led to a relatively diverse set of datasets from different platforms, as discussed in the previous section. This was for example the motivation of [42], who proposed to study personal attacks for the first time on Wikipedia, whereas [43] proposed to look at cyberbullying in short texts, hence focusing on Twitter. Others aimed to focus on cyberbullying events involving media content [38–40], which required exploration of new platforms.

4.2. Session-based data collection

Our exploration of datasets shown in Table 1 shows that 6 of the datasets followed a session-based data collection strategy. This makes it possible to more comprehensively capture the inherent feature of cyberbullying, i.e. repetitive behaviour. It also provides richer data representations that allow models to capture higher-level features. To delve deeper into the use of sessions in cyberbullying datasets, we next discuss datasets created according to two types of sessions: conversation sessions and media sessions.

Conversation sessions. In a typical conversation session, each item presents one question, followed by answers with their associated timestamps [36,37,48]. To fully understand the interactive nature of cyberbullying, the authors of [37] created a MySpace dataset with a moving window to capture each session.

Media sessions. Examples of media-based social networks include Instagram and Vine, where cyberbullying events are perpetrated through media-based communication. [39] collected a large sample of Instagram data, including 3,165,000 media sessions (images and their associated comments) from 25,000 user profiles, of which they labelled a small sample. [40] collected a dataset from the Vine platform, where each post is associated with a video, as well as a collection of likes and comments. While they collected over 650K media sessions, a small sample of it was labelled.

No session-based. The rest of the datasets are not collected based on social media sessions. They are generally made of isolated posts. Authors of [43] collected multiple posts associated with each user timeline, which is different from others collecting individual posts, however, it does not conform to the definition of social media sessions despite being a closer approximation.

4.3. Cyberbullying annotation system

Despite slight variations in wording of the definition of cyberbullying, as well as different interpretations of the overlaps between cyberbullying and related terms (e.g. cyberaggression), there is an overall consistency in referring to the terms “repeatedly”, “intended” and “power imbalance” when defining cyberbullying in the approaches where a rigorous definition is followed, as shown in Table 1.

Following the definition of what constitutes an act of cyberbullying, there have been predominantly two different means for labelling the data. Compared to labelling individual posts, the multimodality of social media sessions makes their labelling particularly challenging [23]. Before the annotation starts, it is important to carefully choose definitions of key terms that will inform annotators during the labelling process. For example, the concept of ‘imbalanced power’ might refer to one user being technically savvier than another, or a popular user abusing less popular users. Repeated cyberbullying can occur over time, such as retweeting/sharing profane comments multiple times [35]. When asking human annotators to determine whether a session constitutes a case of cyberbullying, it is important to incorporate information from all modalities, such as images and text-based comments [23].

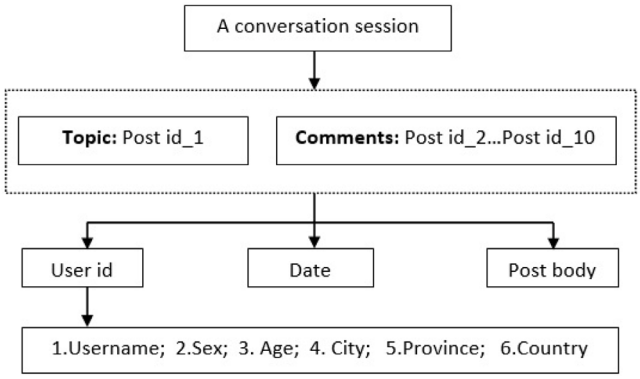


Fig. 5. Illustration of the session structure in the Myspace dataset, where a session takes the form of a conversational thread including a post starting the conversation and followed by others commenting on it.

Crowdsourcing. Most datasets have been annotated by crowd workers through online crowd-sourcing platforms [36,37,39,40,42]. This is indeed a challenging task for crowd workers who are not necessarily trained in identifying cyberbullying, and therefore there is a risk that annotators may end up relying on other factors, such as the use of offensive words, beyond the fact of constituting an act of cyberbullying. To avoid this, [42,43] followed an insightful approach of using previously annotated samples for selecting qualified annotators, i.e. those who got a minimum of labels right would qualify to conduct the rest of the annotation work.

Research assistants. In some cases, researchers have turned to trained annotators to conduct the manual annotation work, as is the case in [37,38]. This has the advantage of having easier interaction with and control of annotators who are known to researchers, as well as the additional advantage of having been trained. It is however costly to recruit expert annotators as well as to train them for the annotation work.

4.4. SSCD datasets

Here we delve into more detail of the three datasets conforming to the two key criteria defined in the SSCD framework, i.e. collected based on social media sessions, and following a rigorous definition of cyberbullying, that incorporates both its repetitive nature and the power imbalance, in addition to providing examples to annotators.

We describe these three datasets next, which were collected from Myspace, Instagram, and Vine:

Myspace [37]. This dataset consists of chat transcripts collected from MySpace.com. These chat transcripts are part of sessions whose structure is illustrated in Fig. 5. These sessions take the shape of conversational threads, where the initial post introduces the discussion topic of the thread, following comments from others in the rest of the thread, which can often end up drifting from the original topic. Within these conversations, each post is considered a constituent part of the session, where a post can be lengthy, e.g. having multiple sentences or paragraphs. Because of the evolving nature of cyberbullying, the conversations were processed using a moving window of 10 posts to capture context. Each post consists of the user profile, date, and content.

Research assistants were provided with detailed annotation guidelines, of which a sample is shown in Fig. 6. In these guidelines, authors divided cyberbullying into 9 categories: flooding, masquerade, flaming, trolling, harassment, cyberstalking and cyberthreats, denigration, outing, and exclusion. Each of the categories is associated with detailed definitions and specific examples. Examples are not simply a single item of cyberbullying, but a cyberbullying session involving multiple user interactions. Two key elements of cyberbullying are taken into account by annotators: recurrence over time and the power of both sides. But clear-cut instruction on how to measure the inequality of power between both parties is lacking. Annotators reviewed each window and were instructed to label whether or not it constituted a case of cyberbullying. Three annotators coded each item, after which the votes were aggregated through majority voting.

Instagram [39]. Instagram is a social media platform where users can post images associated with comments, that others can like or reply to. In this dataset, authors collect each media object and its associated comments, which altogether make a social media session. Each media object contains the following information: media URL, media content, post time, caption, and the number of likes/followed/shared. To facilitate the annotation work, the authors filtered out sessions with fewer than 15 comments. Fig. 7 shows a detailed structure of the sessions as stored in this dataset.

To ensure high-quality annotation, authors restricted annotators to highly rated CrowdFlower workers. In addition, annotators were asked to answer a few test questions prior to taking on the annotation work, to further ensure that annotators were qualified. Annotators were given detailed guidelines including a definition of cyberbullying as well as a set of annotated examples. Annotators would also be disqualified if they completed the annotation work too quickly. In this case, each media session was annotated by five different workers, after which a final label was determined through aggregation. A rigorous definition of cyberbullying that combines negative frequency and power imbalance was used in the labelling process. Power imbalances can take many forms, including physical, social, relational, or psychological, such as one user being more tech-savvy than another, one group of users against one user, or one popular user against one less popular user. Repetition of cyberbullying can happen over time or by retweeting/sharing negative comments or photos by multiple people, increasing the virality of the content. Fig. 8 shows an example of the annotation interface used for this dataset.

Vine [40]. Vine is a video-based online social network, where users can post videos and others can comment on them. This dataset was created by the same authors as the Instagram dataset and therefore followed a very similar approach to collect and annotate this dataset. The dataset in this case is made of media sessions with the same structure as the Instagram data [39], as shown in Fig. 7, with the key difference being that they are initiated by videos rather than images. The annotation methodology is also identical to the Instagram dataset.

5. Recommendations for creating SSCD-based datasets

Existing resources, including both models and datasets, are useful to learn about and design best practices for SSCD dataset creation. Informed by this prior research, we provide recommendations in the three key steps for dataset creation: (i) social media platform selection, (ii) session-based data collection, and (iii) cyberbullying annotation.

5.1. Social media platform selection

The selection of a suitable social media platform for the creation of a dataset should be motivated by the problem and research questions at hand. Where applicable, this motivation can be further strengthened through interdisciplinary collaborations that can help shape stronger and more comprehensive research questions. Still, one of the aspects to take into account is the inherent diversity of cyberbullying and the diverse set of ways in which it is manifested, which also complicates the collection of a dataset encompassing this diversity. Hence, it is also important to clearly define what kinds of cyberbullying events a platform is expected to deliver.

Definition of Flaming: Involves two or more users attacking each other on a personal level. In this form of cyberbullying, there is no clear bully and victim, and there is little to no power differential between the participants. The conversation consists of a heated, short lived argument, and there is bullying language in all of the users' posts.

Example: The conversation consists of a heated, short lived argument, and there is bullying language in all of the users' posts

Person A: "I like hats!"

Person B: "Person A, this is the dumbest blog post on earth. You're the dumbest person on earth. You have no friends. Quit the Internet."

Person A: "Really, well maybe I would, but that would leave only your blog for people to read, and since you have no friends to read it, the Internet would collapse. Fuck your mother."

Person B: "chinga tu madre, coño!" [Translation: "Fuck your mother, mother fucker!"] the participants.

Fig. 6. A fragment of the MySpace annotation guidelines, from [37].

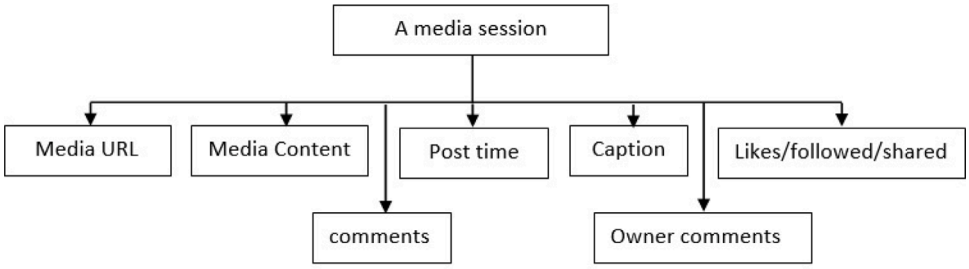


Fig. 7. A media session structure of Instagram and Vine.

5.2. Session-based data collection

The social nature of cyberbullying requires collection beyond simple textual posts, including also its hierarchical structure (i.e. words form comments, comments from conversations), multimodal data (i.e. text, location, user profile, etc.), and evolving user interactions. Hence, datasets should also incorporate this hierarchy, multimodality, and user interactions.

This in turn enables more in-depth and careful implementation of models leveraging the social nature of cyberbullying. For example, session-based investigations can provide valuable insights into the power imbalance between the bully and the victim, which is only likely to manifest across the entire session and may not be observed when looking only at individual texts. The repetitive nature of cyberbullying can be captured by the sequence of comments in a conversation. Examining the hierarchy of social media sessions also enables the model to distinguish the importance of media objects in the session. Thus, session-based detection of cyberbullying opens up promising research directions for identifying, understanding, and ultimately preventing cyberbullying in the real world.

Ensuring that enough cyberbullying cases are captured in a dataset is another challenge, because of the 'rarity' of cyberbullying events if we look at all the content in a social media platform. Many sampling strategies, such as those based on keywords, can introduce a bias in the data selection, and therefore designing a careful data sampling strategy is crucial. At least two promising research directions can help mitigate this bias: (i) intentionally incorporating synthetic yet realistic "perturbations", with the aim of diversifying the content while also preserving its real-world nature, and (ii) careful collection of negative samples, once the positive samples are collected through a carefully designed strategy.

5.3. Cyberbullying annotation

Data annotation is a time-consuming and labour-intensive process. So far, crowdsourcing platforms have been the prevalent option for researchers to annotate datasets. Crowdsourcing has multiple advantages, such as ensuring the diversity and scope of the overall workforce, however it comes at the cost of having a likely untrained set of annotators for what can be considered a relatively challenging annotation work. If crowdsourcing platforms are used, it is advisable to carefully design the annotation guidelines, with sufficient examples including 'edge cases, and to ensure that annotators are qualified, for example through test questions prior to starting the annotation. Previous literature has highlighted the difficulty of distinguishing various types of misconduct [48].

Data labelling is particularly challenging due to the multimodality of social media conversations. When asking annotators to determine whether a conversation constitutes a case of cyberbullying, it is important to integrate all available information with different forms of data components, such as images and text-based comments.

Still, annotation through trained annotators is ideal for a challenging task like cyberbullying, with the main challenges of having access to a set of qualified people, as well as its associated cost.

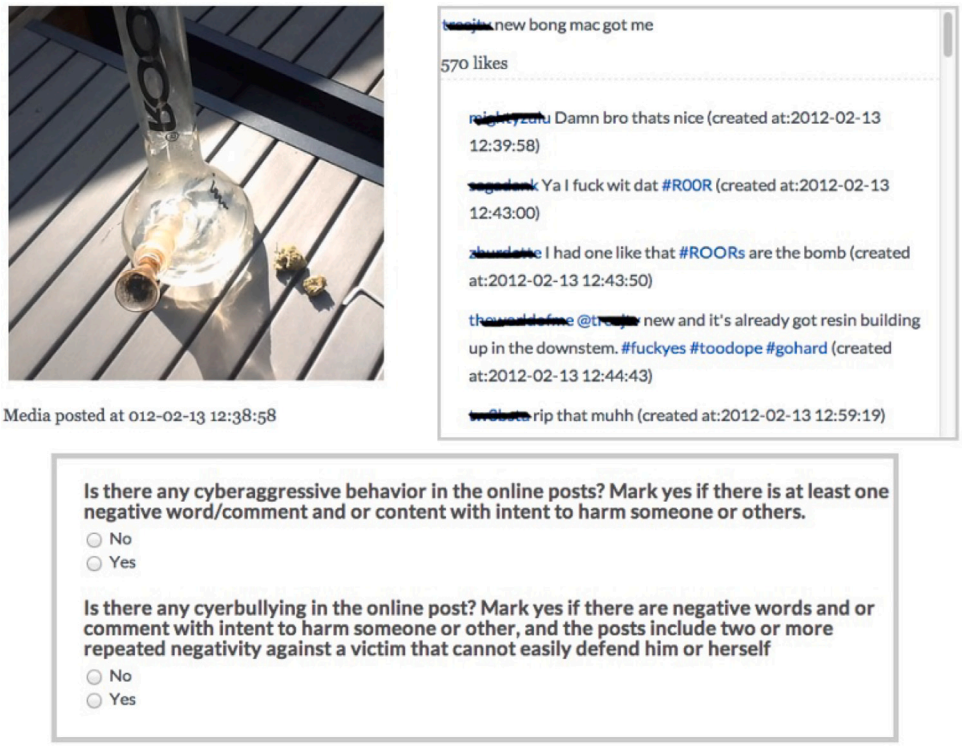


Fig. 8. An example of the labelling study for Instagram datasets. Snapshots are from [39].

6. Trends in cyberbullying detection approaches

In this section, we will discuss general trends in cyberbullying detection approaches, then move on to specific session-based approaches in the next section.

Due to the facts that a large number of methods have been proposed for cyberbullying detection, that these are generally tested on very different environments and settings, and that they are not always reproducible, it is unrealistic to compare them all. Therefore, we focus on analysing the usage trends in types of algorithms, looking at how they are used in different scenarios. Table 2 shows a structured list of the different methods used in cyberbullying detection, showing also the list of research papers where each method has been employed. These methods can be grouped into three types, which we further discuss next: rule-based, machine-learning and deep-learning methods.

6.1. Rule-based methods

Rule-based approaches to cyberbullying detection have been studied for a long period, tracing back to 2008. Then, [49] used subjectivity analysis to design rules for extracting semantic information and keywords for cyberbullying detection. Another influential work using a rule-based approach was proposed by [51]. A framework-based method called lexical syntactic feature (LSF) is proposed to detect offensive content and predict whether a user is a bully, which is determined from a score generated by the model. A sentence-level offensiveness prediction is built, which uses lexical and syntactic features to calculate the offensiveness scores of content, and content-based features and writing style features are adopted to determine user offensiveness scores. Following a similar approach, [52] feeds some person-specific references and multiple curse word dictionaries into a rule-based classifier to tackle the task. Their BullyTracer program used words from the selected dictionary, divided into three categories: insulting words, vulgar language, and pronouns. There has also been researched [70,86–88] analysing the distribution of "Bad words" in corpora, which is then used to identify the most prominent words that help generate a lexicon of "bad words" for cyberbullying detection.

Rule-based approaches have been proven to achieve higher performance in comparison to Naive Bayes and SVM classifiers, but as posited by [51], they are severely limited to the predefined rules and fall short in their ability to generalise to new cases. Moreover, as they do not deal with social media sessions, no further context is used beyond single posts.

6.2. Machine learning methods

Machine learning models are, to date, the most widely used approaches. In addition, hybrid approaches as well as novel machine learning-based frameworks to solve specific complex problems are also being developed.

Off-the-shelf machine learning algorithms. In the early stages of using machine algorithms for cyberbullying detection, much of the research centred around the assessment of which machine learning algorithm performed best for the cyberbullying detection task [36,40,40,54,89] as well as around coming up with effective features to boost model performance through extensive feature engineering [36,43,55–57,61,61,63–66,68,90–92]. Classifiers such as Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), Stochastic Gradient Descent, Bayesian Point Machines, Gradient Boosting, etc. have been extensively tested in various social media platforms such as Twitter, Facebook, MySpace, Formspring.me, Kongregate, Vine, etc. The first traceable work by using a machine learning method can be found in [93], where contextual features were also introduced for the first time for cyberbullying detection. This study builds on the hypothesis that the cyberbullying posts may be short and that detection can be supported through measuring their similarity with neighbouring posts. As well as in many other early cyberbullying detection studies, an SVM classifier is also used in this case.

Use of off-the-shelf machine learning algorithms or proposing improved versions of them is the most widely used research strategy to

Table 2
Summary of cyberbullying detection methods and studies.

Approaches	Methods
Rule-based	1. Semantic based [49,50] 2. Lexical syntactic based [51] 3. Person-specific references and multiple cures word [52]
Machine learning methods	1. Linear/Fuzzy Support Vector Machine [36,38,40,43,53–63] 2. K-nearest neighbours [36,61,63] 3. Logistic Regression [40,43,54,56,64] 4. Naive Bayes multinomial [36,38,40,43,54,61–63,65] 5. Conditional random fields [54] 6. Bayes Point Machine [66] 7. Stochastic Gradient Descent [61] 8. Random Forest [36,38,40,40,43,61,62,65,67–69] 9. Latent Dirichlet Allocation [40,43] 10. Essential Dimensions of Latent Semantic Indexing [70] 11. AdaBoost [40,40] 12. Reinforcement learning [71] 13. Time-Informed Gaussian Mixture Model [72] 14. Fuzzy logic and Genetic algorithm [61,73,74] 15. Participant-Vocabulary Consistency (PVC) using Alternating Least Squares [75] 16. Probabilistic Latent Semantic Analysis [57]
Deep learning	1. Convolutional neural network [67,76–79] 2. (Bidirectional) Long Short-Term Memory [77,78,80] 3. Bidirectional Encoder Representations from Transformers [81] 4. Generative adversarial network [82] 5. Recurrent neural network [77] 6. Semantic-Enhanced Marginalised Denoising Auto-Encoder [83] 7. ConvNet [84] 8. Bi-GRU [85] 9. Hierarchical Attention Network [72]

perform experiments. But with the diversification of social network data types and the in-depth study of cyberbullying detection by researchers, off-the-shelf machine learning algorithms are often found to be limited in dealing with qualitative research of cyberbullying. For example, how to represent multimodal information in social media data (e.g., pictures, videos, user profiles, time and location) [94], and how to model users' characteristics and peer influence [95], or dealing with dynamic data on social platforms [96].

Hybrid approaches. Hybrid techniques combine off-the-shelf machine learning algorithms with a variety of other computational techniques to help improve data analysis and model interpretation. For example, [38,53,73,97] used fuzzy logic to determine the importance scores of different classification models, depending on their advantages. [70] used Latent Semantic Indexing (LSI) [98], a commonly used method to match queries to documents, to match pre-defined cyberbullying query terms with relevant cyberbullying events by using second-order and higher-order word co-occurrences, which helped overcome synonymy and polysemy. [38] adopt an expert system (Multi-Criteria Evaluation Systems (MCES)) with other off-the-shelf machine learning algorithms [59] to extract graph-based features for each post, which is then fed to an SVM classifier. [99] used Latent Dirichlet Allocation (LDA) topic models to predict the probability that a given document belongs to a topic, subsequently using an NB classifier to assign posts to the different categories pertaining to different types of bullies.[58] built a method inspired by the Multiple Sequence Alignment (MSA) method, which is commonly used in computational biology for identifying conserved regions of similarity among raw molecular data. They converted cyberbullying data into string sequences for revealing conserved temporal patterns or slight variations in the attacking strategies of bullies.

Hybrid approaches provide flexibility and transparency, while research proposing novel frameworks enables better adaptation to the task as well as bespoke experimentation and analysis.

Novel frameworks. Given the complexity of cyberbullying datasets, researchers started to develop multi-layer components or multi-model combinations. [60] created a bullying severity identifier composed of multiple fuzzy logic systems, which improve the accuracy of the SVM classifiers to determine the bullying severity through Fuzzy Logic.

Aiming to reduce the number of features used to classify comments and the scalability of online detection, [100] divided the binary classification task into two tasks under a novel framework. One aims to determine if there is an incident of cyberaggression in the comment stream, and the other aims to introduce "repetitiveness" as a threshold to detect session-level cyberbullying. This reduces the risk of large numbers of false positives of past single-text classifiers due to the repetitive nature of cyberbullying.

Feature engineering. Almost all of the supervised learning approaches go through careful feature engineering. In Table 3, we list the set of features that have been used across different studies, along with their associated count. We can see that the most popular features are the *content-based* ones, such as cyberbullying keywords from lexica, topic-based profanity, pronouns, n-grams, Bags-of-words (BOW), Term Frequency Inverse Document Frequency (TF–IDF) etc. Especially, the profanity lexicon is widely used as a cue to detect potential cyberbullying events. However, researchers have pointed out that solely using content-based features can be limited in capturing other inherent characteristics of cyberbullying such as personalisation, contextualisation and diversity, which motivated the use of other features [86].

Sentiment, social and writing style features are also widely used, whereas media-based and demographic features are rarer:

Sentiment features. Most researchers generally use the phrase, keywords and symbols as an indicator of the sentimental expression in a post [35,101–105]. While sentiment features are popular in cyberbullying, they also tend to be insufficient to be used alone and are generally used jointly with other features.

Demographic features. Including the use of gender-specific, age-specific or location vocabularies. [8] also noted that features inferred from author profiles can be effectively used to improve performance.

Social features. Which include features such as followers or online time, tend to be specific to each social media platform and hence more difficult to generalise across platforms, however have also proven to be effective in boosting the performance in specific environments [9,58].

Table 3
List and paper count for different features used in cyberbullying detection.

Type of features	Details	Number of papers
Content features	1. Profanity	23
	2. N-grams	11
	3. Pronouns	10
	4. Cyberbully keywords	8
	5. TF–IDF	7
	6. BOW	3
	7. Skip grams	1
Sentiment features	8. Dictionary of words with sentiment	13
Social features	9. Number comments	3
	10. Number of subscriptions	2
	11. Number of uploads	1
	12. Number of followers	2
	13. Online time	1
	14. Number of friends	1
	15. Ego network	1
Media features	16. ImageNet label	1
Writing style features	17. Length of messages	3
	18. Count/ratio of emoticons	2
	19. Spelling	2
	20. Capitalisation	2
	21. Parts-of-speech tagging	1
	22. The length of the text	1
	23. Number of pronouns	1
Demographic features	24. Age	5
	25. Gender	3
	26. Location	1

Writing-style features. Including features such as "pronoun + profanity" [101], document length [9], word capitalisation [53] and spelling [8], which have shown to be good predictors of a user's likelihood of engaging in abusive behaviour in social media [93].

Media features. As a rather unique and seldom used feature, image-related features were used in the study by [80].

6.3. Deep learning methods

In recent years, there has been a clear shift from the use of machine learning models to an increasing use of deep learning models. With the use of deep learning architectures, other more sophisticated features such as polymorphism, dynamism, hierarchical, and interactivity have also been studied.

Deep learning models have been used to improve representations that are then fed to machine learning algorithms or used in shallow neural networks [69,106–108]. For example, [69] trained a word embedding model that is based on the word2vec skip-gram model for exploring better sentence embeddings, with an RF used for the final classification. Semantic-Enhanced Marginalised Denoising Auto-Encoder [106] (smSDA) was developed via semantic extension of the stacked denoising autoencoder. The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. Linear SVM is then applied to the new feature space.

Deep learning models have also been stacked into hierarchical structures that mirror the complex data structure. [67,71,77–80,84,109,110]. Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Recurrent Neural Networks (RNN) are the most commonly used deep network architectures for these purposes. [67] were the first to use a CNN to transfer from an image classifier to a cyberbullying classifier. Among these deep learning studies, [111] proposed using semantic domain knowledge (demographics, text and social features) to drop out noise and to increase hidden features in the word embedding using stacked deep learning techniques. They then used them in a classification layer for making the whole model more discriminative. [77] analyse cyberbullying detection on various topics across multiple social media platforms using a deep learning model with transfer learning. [112] proposed a double-balanced framework to tackle two important issues: variant contribution and imbalance datasets. [18] used three different types of word embeddings (Word2vec, GloVe, ELMo) that were tested as inputs and coupled with different deep learning architectures.

7. Session-based (SSCD) approaches

In this section, we discuss cyberbullying detection models that adhere to the SSCD framework, which also facilitates a more direct comparison between approaches.

Existing efforts extend text-based analysis to session-based analysis, the extension that is based on the inherent hierarchy of conversations (e.g. word forming comments, dialogue comments), multimodal data (e.g. text, location, images, etc.), and user interactions. To the best of our knowledge, these modelling approaches all emphasise improving the performance of classification tasks by reflecting the nature of cyberbullying. More detailed fine-grained cyberbullying detection, such as how many attacks can be captured in a session and/or at which points of the session, and how to quantify the power of both sides, have not been explored to date.

7.1. Inherent hierarchies with attention

Emerging literature identifies cyberbullying as repetitive temporal acts rather than one-off incidents. Thus, modelling the hierarchical structure of social media sessions and the temporal dynamics of cyberbullying in online social network sessions are key distinctive characteristics of this approach, which are generally considered through three different yet complementary means:

- The hierarchical network structure is adapted to reflect the structure of a social media session.
- Instead of relying on handcrafted features, they leverage an attention mechanism to automatically capture word-level and sentence-level hidden embeddings. They then weight them to form a more representative document-level representation.

- The interval of time between two adjacent comments is considered in a hierarchical network.

HANCD [78] and HENIN [113] can be viewed as two presentive instantiations of this approach. HANCD consists of two levels of Hierarchical Attention Network (HAN): one at the word level and the other at the comment level. These two HANs can capture the differential importance of words and comments in different contexts. Then the bidirectional GRU is employed to capture the sequence of contents. HENIN focuses more on learning various interactions between heterogeneous objects displayed in social media sessions. A comment encoder is created to learn the representations of user comments through a hierarchical self-attention neural network so that the semantic and syntactic cues of cyberbullying can be captured. A post-comment co-attention mechanism learns the interactions between a posted text and its comments. Moreover, two graph convolutional networks are leveraged to learn the latent representations depicting how users interact with each other in sessions, and how posts resemble each other in terms of content.

7.2. Multimodal models

Social media sessions are often multimodal (e.g., image, video, comments, time). Hence, there has also been research in making the most of this diversity of modalities. [114] used encoder denoising techniques and constraints on sparse hidden features. Regarding the method of integrating presentations, a straightforward approach to encode multimodal context is to simply concatenate the raw feature vectors of each modality (e.g., locations, comments, images, timestamps) [92]. However, this method overlooks both structural dependencies among different social media sessions and cross-modal correlations among different modalities. MMCD [92] proposed to train a model based on three different components: (i) Topic-oriented bidirectional long-short term memory (BiLSTM) model with self-attention, (ii) comment-based Hierarchical Attention Network(HAN) to focus on word-level and comments-level characteristics, and (iii) visual embeddings to encode different types of modes. Then they integrated them into a hierarchical attention network to capture hierarchical relationships. XBully [94] is another presentive model, which reformulates multimodal social media data as a heterogeneous network and then aims to learn node embedding representations upon it. In contrast to simply concatenating the raw multi-modal feature vectors of each modality, multiple learned nodes embedded into the resultant heterogeneous network may generate more complex and specific presentations.

7.3. User interaction extractors

Cyberbullying often takes place throughout a series of interactions on social media platforms. Therefore, approaches incorporating sequences of user interactions have also been studied. For example, [50] use a rule-based classifier to tag a conversation session into a sequence of sentiment words to reflect user interactions. [100] proposed a novel algorithm called CONcISE, which reduces the number of classification features used for detecting cyberbullying. The main idea is to feed the sequential aggression detection results of each session into the next high-level cyberbullying detection classifier. Through the so-called Time-Informed Gaussian Mixture Model (UCD), [112] proposed an Unsupervised Cyberbullying Detection method, which incorporates comment inter-arrival times for social media sessions, allowing the use of the full comment history to classify instances of cyberbullying. [115] used a graph neural network for modelling topic coherence and temporal user interactions to capture the repetitive characteristics of bullying behaviour, thus leading to better-predicting performance.

7.4. Performance summary

Most of the SSCD modelling methods mentioned above have experimented on two SSCD datasets: Instagram and Vine. In this section, we summarise and compare the performance of these state-of-the-art methods as shown in their paper. The consistency of how these models have been evaluated facilitates comparison between model performances, which we show in Table 4. Still, it is worth noting that, in addition to the differences in the proposed models, there may be other differences in the preprocessing of the data.

In the comparison of the six models corresponding to the three modelling methods, none of the models achieves consistently the best performance across the two datasets. XBully achieves the best performance among the six models on the Instagram dataset, and MMCD outperforms four other models on the Vine dataset. These two models both adopted multimodal modelling strategies, which suggests that they are promising methods for SSCD. Another interesting observation we make is that the overall performances on Vine are lower than on Instagram, even if the session structure of both datasets is the same.

8. Benchmark experiments with pre-trained language models

In this section, we focus on experimenting and benchmarking the effectiveness of a range of models. While not all the SSCD models presented in Section 7 are available for reproducibility, we present results for two of them: MMCD and XBully. In addition, we test a range of large pre-trained language models: BERT [116], ROBERTA [117], MPNET [118], LONGFORMER [119], XLNet [120], DISTILBERT [121], T5 [122],BERTWEET [123] and ELECTRA [124]. We test all these models on two datasets: Instagram and Vine.

To set up these experiments, we follow the same preprocessing method as [115]. For the implementation of pre-trained language models, we use HuggingFace. The number of training epochs used is 5. We split the data in stratified samples of 80% and 20% for training and testing.

Table 5 shows the Macro-F1 scores of all the models tested. We observe that both MMCD and XBully are competitive models outperforming all pre-trained language models on the Instagram dataset. MMCD and Xbully both belong to the category of multimodal models. In addition to text, data such as time, location, video, and pictures are partially input into the model through embeddings, indicating that multi-modality could help the model to be better understood. However, on the Vine dataset, the majority of the pre-trained models, except for BERT and BERTWEET, outperform both MMCD and XBully. According to a recent study [125] on the Instagram and Vine datasets, it was observed that most cyberbullying incidents happen at the beginning of the Instagram dataset, but these are evenly distributed across sessions in the Vine dataset. MMCD and Xbully use a text truncation strategy when processing long sessions, setting the session length to 140, resulting in a high probability of cyberbullying events being removed after the text truncation in the Vine dataset, while the pre-trained model can accept up to 512 tokens only. If we look at the average performances across both datasets, four pre-trained models, namely ROBERTA, MPNETM, T5 and ELECTRA, show better generalisability than MMCD and XBully.

These experimental results demonstrate that pre-trained language models can be strong, competitive models for Social Media Session-Based Cyberbullying Detection. Still, the differences in performance we observe across both these datasets call for the implementation of more generalisable models that can perform well across different platforms and datasets. This in turn requires the creation and release of additional datasets, ideally from different social media platforms, to further study the generalisability of models beyond these two platforms.

Table 4
Performance comparison of SSCD models on two SSCD datasets: Instagram and Vine.

Approach	Model	Instagram	Vine
Inherent hierarchies with attention	HANCD [78]	0.851	N/A
	HENIN [113]	0.838	0.676
Multimodal model	LSTM + context2vec features [80]	0.85	N/A
	MMCD [114]	0.86	0.841
	XBully [94]	0.878	0.804
User Interaction Extractors	TGBully [115]	0.81 ± 0.02	0.69 ± 0.02

Table 5
Performance of pre-trained language models and state-of-the-art SSCD models.

Model	Instagram	Vine	Average
MMCD	0.86	0.84	0.85
XBully	0.88	0.80	0.84
BERT [116]	0.77	0.83	0.80
ROBERTA [117]	0.85	**0.89**	**0.87**
MPNET [118]	0.85	**0.87**	**0.86**
LONGFORMER [119]	0.77	**0.86**	0.82
T5 [122]	0.79	**0.94**	**0.87**
XLNET [120]	0.83	**0.87**	0.85
ELECTRA [124]	0.83	**0.88**	**0.86**
DISTILBERT [121]	0.82	**0.87**	0.85
BERTWEET [123]	0.76	0.43	0.60

9. Open challenges and conclusion

9.1. Open challenges

Social media session-based cyberbullying detection presents multiple challenges and promising opportunities that differ from single-text-based cyberbullying detection tasks. In this section, we will highlight three open challenges that emerge from our investigation of the subject, related to datasets and models:

Improving the quality of datasets and the clarity in reporting about them. It is often the case that not enough information is reported on how datasets have been created, and how the different underlying factors (i.e. repetition and power imbalance) have been considered if they have. Dictionaries of "bad words" are often used for the data collection, which enables the collection of certain types of cyberbullying but misses other cases where those keywords are not present. This in turn limits the generalisability of the models tested on those datasets, and therefore studying improved data collection strategies should be a priority. In creating cyberbullying datasets, researchers should also avoid conflation with the related concepts of toxicity and hate speech, which differ for example in the fact that they are not necessarily repetitive.

Improving the capacity for fine-grained detection. Since cyberbullying tends to be implicit and subtle in nature, research into more fine-grained detection can be very useful to better understand the phenomenon by pinpointing where it is exactly happening. Thus allowing insight into various characteristics unique to cyberbullying. For example, a detection model should be able to detect not only that social media sessions contain cyberbullying incidents, but also the time period and the number of occurrences. Quantitative indicators of power on both sides.

Increasing the reliability and reproducibility of models. Not all cyberbullying models are reported with sufficient details; where the code of these models is not published, it also means that they are not replicable because the level of detail is insufficient. In order to further research in cyberbullying detection, it is crucial to enable the reproducibility of existing models, so that researchers can build upon and improve existing models. Likewise, it is also important that research in cyberbullying detection considers more than a single dataset in their studies, which enables evaluating the generalisability of models demonstrating competitive performance not only on a single dataset.

9.2. Conclusion

In this survey paper, we review existing approaches to cyberbullying detection, with a particular focus on session-based cyberbullying detection, for which we define the Social media Session-based Cyberbullying Detection framework (SSCD) made of four key components. By going through the research challenges and progress on the four components of the SSCD framework, we review existing research in cyberbullying detection through model and dataset creation, particularly delving into those dealing with social media sessions. In addition, we present a set of comparative, benchmark experiments to evaluate state-of-the-art models on SSCD datasets, as well as posit a set of suggestions for future research when it comes to dataset and model creation.

Through our review, we also highlight the importance of considering two of the inherent characteristics of cyberbullying when designing models, dataset creation, and experiments, i.e. repetition and power imbalance. However, in existing research, these two characteristics have been primarily considered in the annotation stage. Further consideration of these characteristics in the design of cyberbullying detection models is still in its infancy, with a dearth of approaches that incorporate them into the model design. Research into more fine-grained detection will be very useful to better understand the nature of cyberbullying and to advance research in the field.

Where SSCD is an emerging research trend, our survey provides a valuable reference for those studying the problem.

CRediT authorship contribution statement

Peiling Yi: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization, Investigation. **Arkaitz Zubiaga:** Conceptualization, Supervision, Review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Only third-party datasets have been used'.

References

[1] D. Olweus, Bullying at School: What We Know and What We Can Do, Blackwell Publishing, 1993.
[2] P.K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, N. Tippett, Cyberbullying: Its nature and impact in secondary school pupils, J. Child Psychol. Psychiatry 49 (4) (2008) 376–385.
[3] stopbullying, What Is Cyberbullying, URL https://www.stopbullying.gov/cyberbullying/what-is-it.
[4] lawstuff, Cyberbullying, URL https://lawstuff.org.uk/online-safety/cyberbullying/.

[5] R. Broll, L. Huey, "Just being mean to somebody isn't a police matter": Police perspectives on policing cyberbullying, J. School Violence 14 (2) (2015) 155–176.
[6] H. Vandebosch, L. Beirens, W. D'Haese, D. Wegge, S. Pabian, Police actions with regard to cyberbullying: The Belgian case, Psicothema 24 (4) (2012) 646–652.
[7] C.D. Marcum, G.E. Higgins, Examining the effectiveness of academic scholarship on the fight against cyberbullying and cyberstalking, Am. J. Crim. Justice 44 (4) (2019) 645–655.
[8] M. Dadvar, R. Ordelman, F. De Jong, D. Trieschnigg, Improved cyberbullying detection using gender information, in: Dutch-Belgian Information Retrieval Workshop, DIR 2012, 2012, pp. 23–26, URL http://purl.utwente.nl/publications/79872.
[9] M. Dadvar, D. Trieschnigg, F. De Jong, Expert knowledge for automatic detection of bullies in social networks, in: Belgian/Netherlands Artificial Intelligence Conference, 2013, pp. 57–63.
[10] Q. Huang, V.K. Singh, P.K. Atrey, Cyber bullying detection using social and textual analysis, in: SAM 2014 - Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, Workshop of MM 2014, Association for Computing Machinery, Inc, New York, New York, USA, 2014, pp. 3–6, http://dx.doi.org/10.1145/2661126.2661133, URL http://dl.acm.org/citation.cfm?doid=2661126.2661133.
[11] D. Menin, A. Guarini, C. Mameli, G. Skrzypiec, A. Brighi, Was that (cyber)bullying? Investigating the operational definitions of bullying and cyberbullying from adolescents' perspective, Int. J. Clin. Health Psychol. 21 (2) (2021) 100221, http://dx.doi.org/10.1016/j.ijchp.2021.100221, URL https://www.sciencedirect.com/science/article/pii/S1697260021000028.
[12] S.P. Limber, R.M. Kowalski, P.W. Agatston, Cyber Bullying: A Prevention Curriculum for Grades 6-12, Hazelden Publishing, 2008.
[13] J.W. Patchin, S. Hinduja, Cyberbullying Prevention and Response: Expert Perspectives, Routledge, 2012.
[14] R.M. Kowalski, S.P. Limber, P.W. Agatston, Cyberbullying: Bullying in the Digital Age, John Wiley & Sons, 2012.
[15] R.M. Kowalski, G.W. Giumetti, A.N. Schroeder, M.R. Lattanner, Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth, Psychol. Bull. 140 (4) (2014) 1073.
[16] F. Sticca, S. Perren, Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying, J. Youth Adolesc. 42 (5) (2013) 739–750.
[17] K. Van Royen, K. Poels, H. Vandebosch, P. Adam, "Thinking before posting?" Reducing cyber harassment on social networking sites through a reflective message, Comput. Hum. Behav. 66 (2017) 345–352, http://dx.doi.org/10.1016/j.chb.2016.09.040.
[18] H. Rosa, P. Calado, R. Ribeiro, J.P. Carvalho, B. Martins, L. Coheur, Using fuzzy fingerprints for cyberbullying detection in social networks, in: IEEE International Conference on Fuzzy Systems, Vol. 2018-July, Institute of Electrical and Electronics Engineers Inc., 2018, http://dx.doi.org/10.1109/Fuzz-Ieee.2018.8491557.
[19] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ Comput. Sci. 7 (2021) e598.
[20] C. Ziems, Y. Vigfusson, F. Morstatter, Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification, in: Proceedings of the 14th International AAAI Conference on Web and Social Media, No. Icwsm, ICWSM 2020, 2020, pp. 808–819, arXiv:2004.01820.
[21] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, V. Hoste, Automatic detection and prevention of cyberbullying, in: International Conference on Human and Social Analytics, No. c, HUSO 2015, 2015, pp. 13–18, URL https://biblio.ugent.be/publication/7010768/file/7010781.pdf.
[22] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, V. Hoste, Automatic detection of cyberbullying in social media text, 2018, arXiv preprint abs/1801.05617.
[23] L. Cheng, Y.N. Silva, D. Hall, H. Liu, Session-based cyberbullying detection: Problems and challenges, IEEE Internet Comput. 25 (2) (2020) 66–72.
[24] N.S. Ansary, Cyberbullying: Concepts, theories, and correlates informing evidence-based best practices for prevention, Aggress. Violent Behav. 50 (2020) 101343, http://dx.doi.org/10.1016/j.avb.2019.101343, URL https://www.sciencedirect.com/science/article/pii/S1359178918302878.
[25] M. Arif, A systematic review of machine learning algorithms in cyberbullying detection: future directions and challenges, J. Inf. Secur. Cyber. Res. 4 (1) (2021) 01–26.
[26] A. Muneer, S.M. Fati, A comparative analysis of machine learning techniques for cyberbullying detection on Twitter, Future Internet 12 (11) (2020) 187.
[27] A. Kumar, N. Sachdeva, Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis, Multimedia Tools Appl. 78 (17) (2019) 23973–24010.
[28] M.A. Al-Garadi, M.R. Hussain, N. Khan, G. Murtaza, H.F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H.A. Khattak, A. Gani, Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges, IEEE Access 7 (2019) 70701–70718.
[29] H. Rosa, N. Pereira, R. Ribeiro, P. Ferreira, J. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. Veiga Simão, I. Trancoso, Automatic cyberbullying detection: A systematic review, Comput. Hum. Behav. 93 (2019) 333–345.
[30] M. Mladenović, V. Ošmjanski, S.V. Stanković, Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges, ACM Comput. Surv. 54 (1) (2021).
[31] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, Int. J. Surg. 88 (2021) 105906.
[32] E.M. Bender, B. Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Trans. Assoc. Comput. Linguist. 6 (2018) 587–604, http://dx.doi.org/10.1162/tacl_a_00041, URL https://aclanthology.org/Q18-1041.
[33] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering–a systematic literature review, Inf. Softw. Technol. 51 (1) (2009) 7–15.
[34] E.M. Bender, B. Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Trans. Assoc. Comput. Linguist. 6 (2018) 587–604, http://dx.doi.org/10.1162/tacl_a_00041, URL https://aclanthology.org/Q18-1041.
[35] R.I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, S.A. Mattson, Careful what you share in six seconds: Detecting cyberbullying instances in vine, in: J. Pei, F. Silvestri, J. Tang (Eds.), Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, Paris, France, August 25 - 28, 2015, ACM, 2015, pp. 617–622, http://dx.doi.org/10.1145/2808797.2809381.
[36] K. Reynolds, A. Kontostathis, L. Edwards, Using machine learning to detect cyberbullying, in: 2011 10th International Conference on Machine Learning and Applications and Workshops, Vol. 2, IEEE, 2011, pp. 241–244.
[37] J. Bayzick, A. Kontostathis, L. Edwards, Detecting the presence of cyberbullying using computer software, Ursinus College, 2011.
[38] M. Dadvar, D. Trieschnigg, F.d. Jong, Experts and machines against bullies: A hybrid approach to detect cyberbullies, in: Canadian Conference on Artificial Intelligence, Springer, 2014, pp. 275–281.
[39] H. Hosseinmardi, S.A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, S. Mishra, Analyzing labeled cyberbullying incidents on the instagram social network, in: International Conference on Social Informatics, Springer, 2015, pp. 49–66.
[40] R.I. Rafiq, H. Hosseinmardi, S.A. Mattson, R. Han, Q. Lv, S. Mishra, Analysis and detection of labeled cyberbullying instances in vine, a video-based social network, Soc. Netw. Anal. Min. 6 (1) (2016) 1–16, http://dx.doi.org/10.1007/s13278-016-0398-x, URL https://link.springer.com/article/10.1007/s13278-016-0398-x.
[41] J. Sui, Understanding and Fighting Bullying with Machine Learning (Ph.D. thesis), The University of Wisconsin-Madison, 2015.
[42] E. Wulczyn, N. Thain, L. Dixon, Ex machina: Personal attacks seen at scale, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017, ACM, 2017, pp. 1391–1399, http://dx.doi.org/10.1145/3038912.3052591.
[43] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, A. Vakali, Mean birds: Detecting aggression and bullying on twitter, in: Proceedings of the 2017 ACM on Web Science Conference, 2017, pp. 13–22.
[44] J. Wang, K. Fu, C.-T. Lu, Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection, in: 2020 IEEE International Conference on Big Data, Big Data, IEEE, 2020, pp. 1699–1708.
[45] N.V. Chawla, Data mining for imbalanced datasets: An overview, in: Data Mining and Knowledge Discovery Handbook, Springer, 2009, pp. 875–886.
[46] M.A. Al-Garadi, K.D. Varathan, S.D. Ravana, Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network, Comput. Hum. Behav. 63 (2016) 433–443.
[47] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J.P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, E. Dillon, Cyberbullying detection with a pronunciation based convolutional neural network, in: 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2017, pp. 740–745, http://dx.doi.org/10.1109/icmla.2016.0132.
[48] H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, A. Ghasemianlangroodi, Towards understanding cyberbullying behavior in a semi-anonymous social network, in: X. Wu, M. Ester, G. Xu (Eds.), 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, Beijing, China, August 17-20, 2014, IEEE Computer Society, 2014, pp. 244–252, http://dx.doi.org/10.1109/ASONAM.2014.6921591.

[49] A. Mahmud, K.Z. Ahmed, M. Khan, Detecting Flames and Insults in Text, Tech. Rep., BRAC University, 2008, URL http://www.wikipedia.org.

[50] F.K. Ventirozos, I. Varlamis, G. Tsatsaronis, Detecting aggressive behavior in discussion threads using text mining, in: International Conference on Computational Linguistics and Intelligent Text Processing, Springer, 2017, pp. 420–431.

[51] Y. Chen, Y. Zhou, S. Zhu, 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, IEEE, 2012, pp. 71–80.

[52] U. Bretschneider, T. Wöhner, R. Peters, Detecting online harassment in social networks, in: 35th International Conference on Information Systems "Building A Better World Through Information Systems", No. Li 2007, ICIS 2014, 2014, pp. 1–14.

[53] V. Nahar, S. Al-Maskari, X. Li, C. Pang, Semi-supervised learning for cyberbullying detection in social networks, in: Australasian Database Conference, Springer, 2014, pp. 160–171.

[54] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmore, Learning from bullying traces in social media, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 656–666, URL https://aclanthology.org/N12-1084.

[55] S. Parime, V. Suri, Cyberbullying detection and prevention: Data mining and psychological perspective, in: 2014 International Conference on Circuits, Power and Computing Technologies, ICCPCT-2014, IEEE, 2014, pp. 1541–1547.

[56] V.S. Chavan, S. Shylaja, Machine learning approach for detection of cyber-aggressive comments by peers on social media network, in: 2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI, IEEE, 2015, pp. 2354–2358.

[57] R. Zhao, A. Zhou, K. Mao, Automatic detection of cyberbullying on social networks based on bullying features, in: ACM International Conference Proceeding Series, Vol. 04-07-Janu, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1–6, http://dx.doi.org/10.1145/2833312.2849567, URL https://dl.acm.org/doi/10.1145/2833312.2849567.

[58] N. Potha, M. Maragoudakis, D. Lyras, A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data, Knowl.-Based Syst. 96 (2016) 134–155.

[59] E. Papegnies, V. Labatut, R. Dufour, G. Linares, Graph-based features for automatic online abuse detection, in: International Conference on Statistical Language and Speech Processing, Springer, 2017, pp. 70–81.

[60] C.R. Sedano, E.L. Ursini, P.S. Martins, A bullying-severity identifier framework based on machine learning and fuzzy logic, in: International Conference on Artificial Intelligence and Soft Computing, Springer, 2017, pp. 315–324.

[61] B. Haidar, M. Chamoun, A. Serhrouchni, Multilingual cyberbullying detection system: Detecting cyberbullying in arabic content, in: 2017 1st Cyber Security in Networking Conference, CSNet, IEEE, 2017, pp. 1–8.

[62] M.M. Islam, M.A. Uddin, L. Islam, A. Akter, S. Sharmin, 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, IEEE, 2020, pp. 1–6.

[63] D. Soni, V.K. Singh, See no evil, hear no evil: Audio-visual-textual cyberbullying detection, in: Proceedings of the ACM on Human-Computer Interaction, Vol. 2, No. CSCW, ACM New York, NY, USA, 2018, pp. 1–26.

[64] H. Hosseinmardi, R.I. Rafiq, R. Han, Q. Lv, S. Mishra, Prediction of cyberbullying incidents in a media-based social network, in: R. Kumar, J. Caverlee, H. Tong (Eds.), 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016, IEEE Computer Society, 2016, pp. 186–192, http://dx.doi.org/10.1109/ASONAM.2016.7752233.

[65] V. Balakrishnan, S. Khan, H.R. Arabnia, Improving cyberbullying detection using Twitter users' psychological features and machine learning, Comput. Secur. 90 (2020) 101710.

[66] M. Dadvar, D. Trieschnigg, R. Ordelman, F.d. Jong, Improving cyberbullying detection with user context, in: European Conference on Information Retrieval, Springer, 2013, pp. 693–696.

[67] D. Gordeev, Detecting state of aggression in sentences using CNN, in: International Conference on Speech and Computer, Springer, 2016, pp. 240–245.

[68] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, A. Vakali, Detecting aggressors and bullies on Twitter, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 767–768.

[69] T. Bin Abdur Rakib, L.-K. Soon, Using the reddit corpus for cyberbully detection, in: Asian Conference on Intelligent Information and Database Systems, Springer, 2018, pp. 180–189.

[70] A. Kontostathis, K. Reynolds, A. Garron, L. Edwards, Detecting cyberbullying: Query terms and techniques, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci'13, 2013, pp. 195–204, http://dx.doi.org/10.1145/2464464.2464499.

[71] L. Cheng, A. Mosallanezhad, Y. Silva, D. Hall, H. Liu, Mitigating bias in session-based cyberbullying detection: A non-compromising approach, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 2158–2168, http://dx.doi.org/10.18653/v1/2021.acl-long.168, URL https://aclanthology.org/2021.acl-long.168.

[72] L. Cheng, K. Shu, S. Wu, Y.N. Silva, D.L. Hall, H. Liu, Unsupervised cyberbullying detection via time-informed Gaussian mixture model, in: M. d'Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, 2020, pp. 185–194, http://dx.doi.org/10.1145/3340531.3411934.

[73] J. Sheeba, 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, IEEE, 2013, pp. 1–5.

[74] D. Michalopoulos, I. Mavridis, M. Jankovic, GARS: Real-time system for identification, assessment and control of cyber grooming attacks, Comput. Secur. 42 (2014) 177–190.

[75] E. Raisi, B. Huang, Cyberbullying detection with weakly supervised machine learning, in: J. Diesner, E. Ferrari, G. Xu (Eds.), Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, July 31 - August 03, 2017, ACM, 2017, pp. 409–416, http://dx.doi.org/10.1145/3110025.3110049.

[76] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J.P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, E. Dillon, Cyberbullying detection with a pronunciation based convolutional neural network, in: 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA, IEEE, 2016, pp. 740–745.

[77] S. Agrawal, A. Awekar, Deep learning for detecting cyberbullying across multiple social media platforms, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 10772 LNCS, 2018, pp. 141–153, http://dx.doi.org/10.1007/978-3-319-76941-7_11, arXiv:1801.06482.

[78] L. Cheng, R. Guo, Y.N. Silva, D.L. Hall, H. Liu, Hierarchical attention networks for cyberbullying detection on the instagram social network, in: T.Y. Berger-Wolf, N.V. Chawla (Eds.), Proceedings of the 2019 SIAM International Conference on Data Mining, SDM 2019, Calgary, Alberta, Canada, May 2-4, 2019, SIAM, 2019, pp. 235–243, http://dx.doi.org/10.1137/1.9781611975673.27.

[79] V. Banerjee, J. Telavane, P. Gaikwad, P. Vartak, Detection of cyberbullying using deep neural network, in: 2019 5th International Conference on Advanced Computing & Communication Systems, ICACCS, IEEE, 2019, pp. 604–607.

[80] N. Rezvani, A. Beheshti, 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.

[81] S. Paul, S. Saha, CyberBERT: BERT for cyberbullying identification, Multimedia Syst. (2020) 1–8.

[82] P. Yi, A. Zubiaga, Cyberbullying detection across social media platforms via platform-aware adversarial encoding, in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 16, 2022, pp. 1430–1434.

[83] R. Zhao, K. Mao, Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder, IEEE Trans. Affect. Comput. 8 (3) (2016) 328–339.

[84] A. Kumar, N. Sachdeva, Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network, Multimedia Syst. (2021) 1–10.

[85] L. Cheng, R. Guo, Y.N. Silva, D. Hall, H. Liu, Modeling temporal patterns of cyberbullying detection with hierarchical attention networks, ACM/IMS Trans. Data Sci. 2 (2) (2021) 1–23.

[86] S. Salawu, Y. He, J. Lumsden, Approaches to automated detection of cyberbullying: A survey, IEEE Trans. Affect. Comput. 11 (1) (2020) 3–24, http://dx.doi.org/10.1109/TAFFC.2017.2761757.

[87] M. Mladenović, V. Ošmjanski, S.V. Stanković, Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges, ACM Comput. Surv. 54 (1) (2021) 1–42.

[88] C. Emmery, B. Verhoeven, G. De Pauw, G. Jacobs, C. Van Hee, E. Lefever, B. Desmet, V. Hoste, W. Daelemans, Current limitations in cyberbullying detection: on evaluation criteria, reproducibility, and data scarcity, 2019, arXiv preprint abs/1910.11922.

[89] K. Balci, A.A. Salah, Automatic analysis and identification of verbal aggression and abusive behaviors for online social games, Comput. Hum. Behav. 53 (2015) 517–526.
[90] V. Balakrishnan, S. Khan, T. Fernandez, H.R. Arabnia, Cyberbullying detection on twitter using big five and dark triad features, Pers. Individ. Differ. 141 (2019) 252–257.
[91] A. Bozyiğit, S. Utku, E. Nasibov, Cyberbullying detection: Utilizing social media features, Expert Syst. Appl. 179 (2021) 115001.
[92] N. Potha, M. Maragoudakis, Cyberbullying detection using time series modeling, in: 2014 IEEE International Conference on Data Mining Workshop, IEEE, 2014, pp. 373–382.
[93] D. Yin, Z. Xue, L. Hong, B.D. Davison, A. Kontostathis, L. Edwards, Detection of harassment on web 2.0, in: Proceedings of the Content Analysis in the WEB, Vol. 2, 2009, pp. 1–7.
[94] L. Cheng, J. Li, Y.N. Silva, D.L. Hall, H. Liu, XBully: Cyberbullying detection within a multi-modal context, in: J.S. Culpepper, A. Moffat, P.N. Bennett, K. Lerman (Eds.), Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019, ACM, 2019, pp. 339–347, http://dx.doi.org/10.1145/3289600.3291037.
[95] L. Cheng, J. Li, Y. Silva, D. Hall, H. Liu, PI-bully: Personalized cyberbullying detection with peer influence, in: The 28th International Joint Conference on Artificial Intelligence, IJCAI, 2019.
[96] B.A.H. Murshed, J. Abawajy, S. Mallappa, M.A.N. Saif, H.D.E. Al-Ariki, DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform, IEEE Access 10 (2022) 25857–25871.
[97] B.S. Nandhini, J. Sheeba, Online social network bullying detection using intelligence techniques, Procedia Comput. Sci. 45 (2015) 485–492.
[98] A. Kontostathis, Essential dimensions of latent semantic indexing (LSI), in: 2007 40th Annual Hawaii International Conference on System Sciences, HICSS'07, 2007, p. 73, http://dx.doi.org/10.1109/HICSS.2007.213.
[99] Z. Ashktorab, E. Haber, J. Golbeck, J. Vitak, Beyond cyberbullying: self-disclosure, harm and social support on ASKfm, in: Proceedings of the 2017 ACM on Web Science Conference, 2017, pp. 3–12.
[100] M. Yao, C. Chelmis, D.-S. Zois, Cyberbullying ends here: Towards robust detection of cyberbullying in social media, in: The World Wide Web Conference, 2019, pp. 3427–3433.
[101] K. Dinakar, R. Reichart, H. Lieberman, Modeling the detection of textual cyberbullying, in: AAAI Workshop, Technical Report WS-11-02, 2011, pp. 11–17.
[102] S.M. Serra, 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 - Proceedings of the ISSA 2011 Conference, 2011, http://dx.doi.org/10.1109/ISSA.2011.6027507.
[103] H. Sanchez, S. Kumar, Twitter Bullying Detection, Tech. Rep., 2011, URL https://www.researchgate.net/publication/267823748.
[104] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, R. Picard, Commonsense reasoning for detection, prevention, and mitigation of cyberbullying, in: BodyNets International Conference on Body Area Networks, 2012, http://dx.doi.org/10.1145/0000000.0000000.
[105] G. Fahrnberger, D. Nayak, V.S. Martha, S. Ramaswamy, SafeChat: A tool to shield children's communication from explicit messages, in: 14th International Conference on Innovations for Community Services: "Technologies for Everyone", I4CS 2014 - Conference Proceedings, IEEE, 2014, pp. 80–86, http://dx.doi.org/10.1109/I4CS.2014.6860557, URL http://ieeexplore.ieee.org/document/6860557/.
[106] R. Zhao, K. Mao, Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder, IEEE Trans. Affect. Comput. 8 (3) (2017) 328–339, http://dx.doi.org/10.1109/TAFFC.2016.2531682.
[107] O. Gencoglu, Cyberbullying detection with fairness constraints, IEEE Internet Comput. 25 (1) (2020) 20–29.
[108] S. Pericherla, E. Ilavarasan, Transformer network-based word embeddings approach for autonomous cyberbullying detection, Int. J. Intell. Unmanned Syst. (2021).
[109] S.-J. Bu, S.-B. Cho, A hybrid deep learning system of CNN and LRCN to detect cyberbullying from SNS comments, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, 2018, pp. 561–572.
[110] C. Iwendi, G. Srivastava, S. Khan, P.K.R. Maddikunta, Cyberbullying detection solutions based on deep learning architectures, Multimedia Syst. (2020) 1–14.
[111] R. Zhao, K. Mao, Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder, IEEE Trans. Affect. Comput. 8 (3) (2017) 328–339, http://dx.doi.org/10.1109/TAFFC.2016.2531682.
[112] L. Cheng, R. Guo, K.S. Candan, H. Liu, Representation learning for imbalanced cross-domain classification, in: C. Demeniconi, N.V. Chawla (Eds.), Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020, Cincinnati, Ohio, USA, May 7-9, 2020, SIAM, 2020, pp. 478–486, http://dx.doi.org/10.1137/1.9781611976236.54.
[113] H.-Y. Chen, C.-T. Li, HENIN: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, Online, 2020, pp. 2543–2552, http://dx.doi.org/10.18653/v1/2020.emnlp-main.200, URL https://aclanthology.org/2020.emnlp-main.200.
[114] K. Wang, Q. Xiong, C. Wu, M. Gao, Y. Yu, Multi-modal cyberbullying detection on social networks, in: 2020 International Joint Conference on Neural Networks, IJCNN, 2020, pp. 1–8, http://dx.doi.org/10.1109/IJCNN48605.2020.9206663.
[115] S. Ge, L. Cheng, H. Liu, Improving cyberbullying detection with user interaction, in: Proceedings of the Web Conference 2021, WWW '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 496–506, http://dx.doi.org/10.1145/3442381.3449828.
[116] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.
[117] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint abs/1907.11692.
[118] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Adv. Neural Inf. Process. Syst. 33 (2020) 16857–16867.
[119] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The long-document transformer, 2020, arXiv preprint abs/2004.05150.
[120] Z. Yang, Z. Dai, Y. Yang, J.G. Carbonell, R. Salakhutdinov, Q.V. Le, XLNet: Generalized autoregressive pretraining for language understanding, in: H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 5754–5764, URL https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.
[121] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019, arXiv preprint abs/1910.01108.
[122] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P.J. Liu, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (140) (2020) 1–67.
[123] D.Q. Nguyen, T. Vu, A. Tuan Nguyen, BERTweet: A pre-trained language model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 9–14, http://dx.doi.org/10.18653/v1/2020.emnlp-demos.2, URL https://aclanthology.org/2020.emnlp-demos.2.
[124] K. Clark, M.-T. Luong, Q.V. Le, C.D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, 2020, arXiv preprint arXiv:2003.10555.
[125] P. Yi, A. Zubiaga, Learning like human annotators: Cyberbullying detection in lengthy social media sessions, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 4095–4103.

Peiling Yi, A research student in the Cognitive Science Research Group, Queen Mary University of London. A large part of her research interests falls in the intersection of transfer learning and text classification. Currently, her main research interest is in cyberbullying detection across different social media platforms.

Dr. Arkaitz Zubiaga, Senior Lecturer (Associate Professor) at the Queen Mary University of London, where he leads the Social Data Science lab. His research interests revolve around linking online data with events in the real world, among others for tackling problematic issues on the Web and social media that can have a damaging effect on individuals or society at large, such as hate speech, misinformation, inequality, biases and other forms of online harm.