IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 8, 2021, accepted July 17, 2021, date of publication July 21, 2021, date of current version July 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3098979

When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection

FATMA ELSAFOURY[1,2], STAMOS KATSIGIANNIS[3], (Member, IEEE), ZEESHAN PERVEZ[1], (Senior Member, IEEE), AND NAEEM RAMZAN[1], (Senior Member, IEEE)

[1]School of Computing, Engineering and Physical Sciences, University of the West of Scotland, Paisley PA1 2BE, U.K.
[2]Seric Systems Ltd., Glasgow G2 6TS, U.K.
[3]Department of Computer Science, Durham University, Durham DH1 3LE, U.K.

Corresponding author: Fatma Elsafoury (fatma.elsafoury@uws.ac.uk)

This work was supported by the UK Research and Innovation (UKRI) as a Knowledge Transfer Partnership (KTP) under Project 11465.

ABSTRACT Web 2.0 helped user-generated platforms to spread widely. Unfortunately, it also allowed for cyberbullying to spread. Cyberbullying has negative effects that could lead to cases of depression and low self-esteem. It has become crucial to develop tools for automated cyberbullying detection. The research on developing these tools has been growing over the last decade, especially with the recent advances in machine learning and natural language processing. Given the large body of work on this topic, it is vital to critically review the literature on cyberbullying within the context of these latest advances. In this paper, we survey the automated detection of cyberbullying. Our survey sheds light on some challenges and limitations for the field. The challenges range from defining cyberbullying, data collection, and feature representation to model selection, training, and evaluation. We also provide some suggestions for improving the task of cyberbullying detection. In addition to the survey, we propose to improve the task of cyberbullying detection by addressing some of the raised limitations: 1) Using recent contextual language models like BERT for the detection of cyberbullying; 2) Using slang-based word embeddings to generate better representations of the cyberbullying-related datasets. Our results show that BERT outperforms state-of-the-art cyberbullying detection models and deep learning models. The results also show that deep learning models initialized with slang-based word embeddings outperform deep learning models initialized with traditional word embeddings.

INDEX TERMS Cyberbullying detection, deep learning, social media, text classification.

I. INTRODUCTION

The internet has become an important development tool for young people. It provides a great source of information and a tool for communication. In recent studies, children and young people categorized their Internet activities into three groups: (a) Content-based activities, such as school work, play games, watch video clips, read the news, or download music; (b) Contact/communication-based activities such as instant messaging, email, chatting or Skype; and (c) Conduct peer participation activities such as blogging, post photos or file-sharing websites [1]. Despite all the benefits, the Internet could be an environment for bullying. In their research, Haddon and Livingstone [2] showed that 17% of the children, who were interviewed between the age of 9 and 14 in the UK, were exposed to sexual content compared to 24% of children from the EU. The study also showed that the children experienced bad language in the form of insults or swearing, aggressive communication, or harassment. Moreover, social media platforms provide a fruitful environment for cyberbullying in the forms of threats, harassment, and exploiting potential victims [3]. The Pew research center reported in 2017 that 40% of social media users have experienced some form of cyberbullying [4]. Another study that included university students found that among 200 university students, 91% experienced cyberbullying, 55.5% of them on Instagram, and 38% on Facebook [5].

Cyberbullying experiences can have serious consequences for the victims, including depression, anxiety, low self-esteem, and self-harm, and may even lead in extreme cases

The associate editor coordinating the review of this manuscript and approving it for publication was Nazar Zaki.

TABLE 1. Discussed sections in published literature review papers on the automated detection of cyberbullying.

Paper	Year	Systematic review	Definition section	Types section	Data section	Data annotation section	Features section	Preprocessing section	Models section	Evaluation metrics section	Replication experiments section	Extended experiments section
[24]	2015	✓										
[25]	2016			✓	✓							
[26]	2016		✓	✓					✓	✓		✓
[27]	2017	✓	✓		✓		✓	✓	✓			
[28]	2017											
[29]	2018		✓	✓	✓				✓			
[30]	2018	✓	✓		✓		✓			✓		✓
[31]	2018	✓	✓		✓	✓	✓		✓			
[32]	2018		✓		✓						✓	✓
[33]	2018		✓	✓	✓		✓		✓			
[34]	2018	✓	✓	✓	✓		✓		✓			
[35]	2019	✓	✓				✓		✓	✓		
[36]	2019				✓		✓				✓	✓
[37]	2020	✓	✓	✓	✓	✓						
[38]	2020	✓	✓	✓	✓	✓						
[39]	2021	✓	✓	✓	✓	✓		✓				
This	2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

to suicide [6]. Consequently, having tools for detecting and preventing cyberbullying is crucial for reducing the negative effects. Studying cyberbullying is rooted in Psychology, Education, Behavioural Science (BS), and Information Technology (IT). On the IT front, the automated detection of cyberbullying can help in the automated removal of the flagged content, post, or communication, in the automated blocking of the perpetrators, and in reaching out to help the victims. Over the last decade, the body of literature on automated detection of cyberbullying has been growing, especially on the topic of detecting cyberbullying from social media networks like Twitter [7]–[13], Instagram [10], [11], [14]–[16] and YouTube [17]–[19]. This body of research has been working towards automated cyberbullying detection using either rule-based models [12], [19], [20], conventional machine learning models [16], [18], [19], [21], or deep learning models [13], [21]–[23].

The last decade brought significant advances in the fields of Machine Learning (ML) and Natural Language Processing (NLP), which have been successfully applied in domains related to cyberbullying detection, such as rumor detection [40], sentiment analysis [41], and fake news detection [42]. Consequently, it is extremely useful to review the available literature on automated cyberbullying detection, in light of these recent advances. There have been various attempts to review that body of literature. An overview of the published literature review papers between 2009 and 2021 on the topic of automated cyberbullying detection is provided in Table 1. The works shown in Table 1 cover the following aspects of the examined problem: systematic review or how the literature was collected [27], [31], [32]; cyberbullying definition [32], [33], [35], [39]; cyberbullying types [25], [26], [29]; datasets [25], [28], [32]; feature selection [27], [35], [36]; model selection [26], [27], [31]; and evaluation metrics [26], [30], [35]. However, only few are comprehensive [26], [30]. There are some important aspects that are rarely covered in the literature, like data annotation [31] and data preprocessing [27]. In addition, some review papers replicate experiments from their reviewed literature [31], [36], while others design their own experiments to fill in gaps in the literature [26], [30].

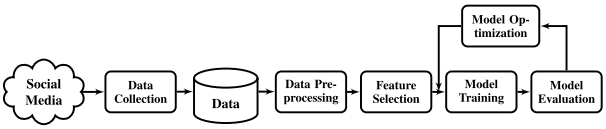


FIGURE 1. Machine learning pipeline.

However, none of the reviews from Table 1 organize the reviewed literature around the steps of the machine learning (ML) pipeline. The ML pipeline (Figure 1) is a series of ordered steps that constitute the machine learning workflow, consisting of data collection (data sourcing and data annotation), data pre-processing, feature selection, model training, and model evaluation [43]. Organizing the literature review around the ML pipeline would help to aggregate the different methods and approaches used to accomplish each step in the pipeline, giving the reader the opportunity to learn and compare these different approaches and methods. Taking this into consideration, in this work, we organized our reviewed literature around the steps of the ML pipeline employed by each reviewed work.

Our literature review on the automated detection of cyberbullying sheds light on some of the challenges and the limitations of the current literature: (a) Data collection; (b) Features selection; (c) Models selection and training; and (d) Evaluation metrics. We also provide some suggestions to address

some of the limitations and to improve the task of cyberbullying detection. Among these suggestions are: i) Using recent contextual language models like the Bidirectional Encoder Representations from Transformers (BERT) and ii) Using slang-based word embeddings as feature representation of the cyberbullying-related datasets.

In addition to reviewing the literature, we investigate the impact of our suggestions on improving the task of cyberbullying detection. We start our investigation by replicating one of the state-of-the-art models in detecting cyberbullying in the reviewed literature and use it on more datasets. Then, to investigate the impact of using contextual language models like BERT to improve the detection of cyberbullying, we compare the performance of BERT, fine-tuned on cyberbullying-related datasets, to the replicated study, and state-of-the-art deep learning models. Then, to find out if using slang-based word embedding would improve the task of the detection of cyberbullying, we ran a series of experiments to compare different deep learning models trained on randomly initialized word embeddings, traditional word embeddings, and slang-based word embeddings.

The contributions of this literature review paper can be summarised as follows:

1) A systematic literature review on automated cyberbullying detection that covers all the steps in the machine learning pipeline.
2) Demonstrating that contextual language models like BERT improve the detection of cyberbullying on the used datasets.
3) Demonstrating that using slang-based word embeddings improves the detection of cyberbullying.

This paper is organized into two parts. In the first part, we reviewed the collected body of literature on automated cyberbullying detection, starting with explaining our search strategy for selecting literature works (Section II) and then reviewing the different definitions of cyberbullying in the literature and the different types of cyberbullying (Section III). Then, we reviewed the different methods used in the literature for each step in the machine learning pipeline: data collection (Section IV-A), pre-processing (Section IV-B), feature selection (Section IV-C), model training (Section IV-D) and model evaluation (Section IV-E). Then, we provide a critical analysis of the current challenges and limitations in the literature of cyberbullying detection (Section V).

In the second part of the paper, we presented our experimental evaluation (Section VI) for the replicated study (Section VI-A); we used the replicated study on other datasets (Section VI-B); we fine-tuned BERT on cyberbullying-related datasets (Section VI-C); and we used slang-based word embeddings (Section VI-D). Finally, we provided an insight into our results, drew our conclusions, and discussed possible future work in Section VII.

II. SEARCH STRATEGY AND STUDY SELECTION

The papers reviewed in this work were selected by following a systematic literature review method to make sure that as many relevant papers as possible were covered. To achieve this, we first looked at how other literature reviews selected their papers. Among the literature review papers in Table 1, the collection methods used in [27] and [31] generated the highest number of relevant papers, which is 43. They used the search keywords "cyberbullying" and "detection" to search through the Google Scholar, IEEE Xplore, Science Direct, ACM Digital Library and Wiley online databases. Following their method, we located some of the key studies in the field of automated cyberbullying detection. To ensure that as many relevant and new papers as possible are covered, we reviewed the papers that cited those key studies and especially those published after 2016. This process led to 106 papers related to computational methods for cyberbullying detection. Figure 2 shows the number of the reviewed papers grouped by the publication year from 2008 to 2020. It must also be noted that the list of the papers reviewed in this work, as well as the Python scripts used to analyse the data and the scripts used for the experimental evaluation, will be made publicly available upon acceptance of this work.

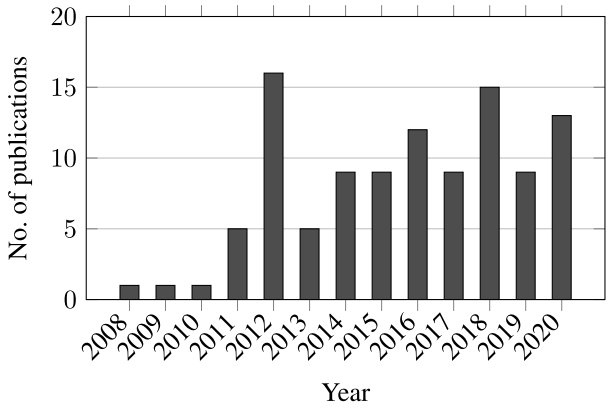


FIGURE 2. The number of papers on automated detection of cyberbullying that we reviewed, grouped by the year of publication from 2008 to 2020.

III. CYBERBULLYING

A. DEFINITION

The lack of a globally accepted definition of cyberbullying is one of the main issues detected in the reviewed literature on automated cyberbullying detection. For example, although some of the reviewed works claim to detect cyberbullying in their title, they detect child grooming [52], [70] or detect the participants in the act, like the bullies, victims, and bystanders, rather than the actual incident of cyberbullying [71], [72]. Out of the 106 reviewed papers, 65 papers defined cyberbullying. There are eight main definitions that most of the papers used, as shown in Table 2. However, despite these definitions being close in meaning, as most of them describe cyberbullying as "*one form or another of insulting, spread using mobile or internet technology*", the lack of a clear definition leads to difficulties in comparing and evaluating different works. For example, in [19], [51], [57], cyberbullying is described as online

TABLE 2. The most common cyberbullying definitions used in the reviewed literature.

Definition	Used in
Cyberbullying is a form of cyber-aggression that is defined as an intentional harmful act to another person that takes place through online means and is characterized by imbalance of power between the individuals involved and repetition of the act [44]–[46]	[47]–[50]
Cyberbullying is an individual's intentional and repeated harmful act to others through harmful posts or messages through various digital technologies [51]	[52]–[56]
The use of electronic forms of communication to abuse, threat, or harass another person [57]	[10], [58]–[60]
When the Internet, cell phones, or other devices are used to send or post text or images intended to hurt or embarrass another person [61]	[52]–[56]
Willful and repeated harm inflicted through the medium of electronic text [62]	[20], [62]–[64]
Online harassment is to include being called offensive names, purposefully embarrassed, stalked, sexually harassed, physical threat in a sustained manner [66]	[66]
Any fierce, purposeful activity directed by people or gatherings, utilizing on the web channels over and again against a victim who does not can [68]	[68], [69]
Hate speech is defined as targeting individuals or groups based on their characteristics (targeting characteristics); demonstrating a clear intention to incite harm, or to promote hatred; it may or may not use offensive or profane words [8]	[8], [13]

aggression, bullying using new communication technologies, online harassment, or hate speech. This is problematic as each of these tasks is different, making it significantly difficult to replicate the studies and to compare the models' results and generalisability. Some studies consider cyberbullying as a sub-type of cyber-aggression [44], while others consider cyberbullying as a different task from cyber-aggression [73]. Mladenović *et al.* provided a detailed survey on the diversity of the definitions of cyberbullying, cyber-aggression, trolling, and cyber-grooming [39]. Another issue is that some studies do not differentiate between bullying and cyberbullying apart from the usage of electronic means. As a consequence, they require the following three characteristics of bullying to be evident in cyberbullying cases: harmful, repetitive, and with power imbalance between the bully and the victim. These characteristics sometimes are hard to satisfy in the online space. For example, someone may send a bullying message to someone during an online conversation only once, which does not satisfy repetition. However, some studies claim that the fact that an online post makes permanent harm satisfies the repetition requirement [31]. In addition, in the case of the Twitter platform, Tian and Xin argue that negative messages on Twitter tend to be retweeted more often, which also satisfies the repetition requirement [74].

B. CYBERBULLYING TYPES

According to the literature, there are 12 types of cyberbullying [29]:

1) Flamming: Starting a fight online.
2) Harassment: Sending insulting messages frequently.
3) Cyberstalking: Sending intimidating messages to the victim, which causes fear.
4) Masquerade: The bully pretends to be someone else.
5) Trolling: Posting controversial comments to upset other members on the online platform.
6) Denigration: Negative gossip about another person.
7) Outing: Posting personal information about someone in public forums.
8) Exclusion: When a social group deliberately excludes someone.
9) Catfishing: Creating a fake profile using someone else's information.
10) Dissing: Posting information about someone to hurt them or defame them.
11) Trickery: Tricking someone to share their secrets or personal information.
12) Fraping: Using someone else's online account to post inappropriate content and tricking others into believing that the account owner posted them.

Most of the reviewed literature does not specify which type of cyberbullying they are detecting. Nevertheless, online harassment is the most common type of cyberbullying in the literature [49], [50], [56], [75]–[77]. There are sub-types of harassment mentioned in the reviewed literature like Aggression [15], [64] and Toxicity [66].

C. HATE SPEECH

In the last few years, research on hate speech detection has been increasing [8], [13], [18], [21], [78], [79]. In a survey paper on the automated detection of hate speech in text, Fortuna and Nunes studied the definition of hate speech in the literature in relation to four dimensions: physical violence encouragement, targets, attack language, and humorous hate speech [34]. From these four dimensions, the authors proposed a new definition for hate speech, i.e. "*Hate speech is a language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used*".

In the NLP community, it is unclear what the difference in definition between hate speech and cyberbullying is. This lack of clarity can cause generalisability problems with the developed models, as each of the cyberbullying detection and hate speech detection tasks require different features. However, there are also some similarities between the two tasks. The main similarity is the abusive language, while the main difference is the target of the abusive language. In cyberbullying, the abusive language is targeted at specific

TABLE 3. Types of hate speech and their targets in the literature [80].

Categories	Example of possible targets
Race	Black people, white people
Behaviour	Insecure people, sensitive people
Physical	Obese people, beautiful people
Sexual orientation	Gay people, straight people
Class	Ghetto people, rich people
Gender	Pregnant people, women
Ethnicity	Chinese people, Indian people
Disability	Bipolar people, people with mental disabilities
Religion	Religious people, Muslims, Jews, Atheists
Other	Drunk people, shallow people

individuals, while hate speech is targeted at groups of people who share specific characteristics [34]. Examples of types of hate speech and their targeted groups are summarised in Table 3.

The main focus of this paper is to review the literature on cyberbullying detection. However, due to the similarities between cyberbullying and hate speech, we opted to include some of the hate speech datasets and features used in the literature in addition to the cyberbullying datasets and features. Consequently, the term cyberbullying will hereby cover both hate speech and cyberbullying in this work.

IV. MACHINE LEARNING PIPELINE

This section provides a thorough literature review on automated cyberbullying detection, organized by the steps in the machine learning pipeline, as shown in Figure 1.

A. DATA COLLECTION

In the reviewed literature, the used datasets originated from various social media platforms. In this section, we provide an overview of the different datasets used in the literature, including the annotation processes followed, the ratio between positive and negative samples, and the sampling strategies used.

1) DATA SOURCES

The datasets used in the reviewed literature originated from twelve different sources, including seven social media platforms (Twitter, Instagram, FormSpring, Ask.FM, MySpace, YouTube, Vine, and Reddit), an online collaborative platform (Wikipedia Talk Pages), and a news website (Yahoo News). All of these platforms have experienced incidents of cyberbullying and were thus used for the creation of datasets for cyberbullying detection. Examples of offensive comments from these data sources can be found in Table 4. In addition, details about all the datasets, including their source, the number of positive and negative samples, the proportion of positive vs. negative samples, their focus (e.g., cyberbullying, hate speech, cyber-aggression, etc.), their availability, and related references, are provided in Table 5.

- **Twitter** is one of the most famous social media platforms where cyberbullying takes place [74]. In the reviewed cyberbullying literature, there are 13 datasets collected from Twitter with different sizes, collection methods, and annotation methods. The tweets in the datasets were collected using the public Twitter API.[1] Some studies used hateful hashtags and profane words, like feminazi, immigrant, nigger, Islam, terrorism, and bully to filter the tweets [7]–[13]. Other studies used publicly available datasets, like for example [90] and [84], who used the 2011 TREC Microblog Track corpus.[2] In 2019, the multilingual detection of hate speech against immigrants and women in Twitter (hateEval)[3] dataset was released in two languages, English and Spanish. The dataset was used in SemEval 2019 Task 5 [86].
- **Instagram**[4] is a social media platform where people share photos and videos, and others can comment on them. This opens the door for cyberbullying, as people can either post offensive pictures or write insulting comments. In the reviewed literature, we found four Instagram datasets [10], [11], [14], [15]. The data was crawled from Instagram by first filtering images and videos using hate speech, harassment, and abusive words. Then, collecting those media sessions where offensive comments were made.
- **FormSpring.ME**[5] is a social media platform that allows its users to ask other users anything and start a conversation between them. Sometimes the questions or the answers are abusive. In the reviewed literature, there are three FormSpring.ME datasets [20], [21], [53]. Two datasets were made available as part of the Kaggle competition website[6] and were used by [21], [53], [55]. The third dataset was crawled from the FormSpring.ME website by [20] and made available by the researchers.[7]
- **Ask.FM**[8] is a social media website that is similar to FormSpring.ME, where users can ask other users questions and start a conversation. In the reviewed literature, we found two studies that used data from ASK.FM [10], [11]. The data was crawled from the ASK.FM website. The researchers used a custom-made harassment dictionary to query data from other sources but it is not clear if they used the same method to filter the crawled data from ASK.FM or not.
- **MySpace**[9] is a social networking website that used to be very famous in the 2000s. In the reviewed literature, two studies used data from MySpace. The dataset was collected by [63] and then was used by [83].

[1] https://developer.twitter.com/en/docs
[2] https://trec.nist.gov/data/microblog2011.html
[3] https://docs.google.com/forms/d/e/1FAIpQLSdLGwaiPd-JhNCfnRQxELp9YVB8GrNFGkQieWqENSruIqMRPw/closedform
[4] https://www.instagram.com/
[5] https://domain.me/formspring-me/
[6] 2https://www.kaggle.com/swetaagrawal/formspring-data-for-cyberbullyingdetection
[7] https://www.chatcoder.com/drupal/DataDownload
[8] https://ask.fm/
[9] https://myspace.com/

TABLE 4. Examples of cyberbullying comments on social media.

Comment	Source	Label
my boyfriend showed this song to me I love it Me tooo Is she having a seizure. Omg u have a corgi I am training for the Olympics and I am Russian You want some rapes... LOL	YouTube	Cyberbullying (Aggression)
RT @BeepsS: @senna1 @BeepsS: I'm not sexist but f**k if you're a woman and you can't Cook get your s**t together.	Twitter	Cyberbullying (Sexist)
@freemedialive F**k #Islam. Mohammed was a pedophile, murderer, bigot, sexist, rapist, slave trader, caravan robber, and liar. : racism	Twitter	Cyberbullying (Racist)
You f**k your dad.	Kaggle-insults	Cyberbullying (Insult)
f**k off you little a**hole. If you want to talk to me as a human start showing some fear the way humans act around other humans, because if you continue your beligerant campaign, i will cross another boundary and begin off-site recruitmehnt. I can escalate till I am rhetorically nuclear with the whole goddamed mob of you if that is where you think you will find what you want. You had better start expressing some interest in the concerns presented to you or your credibility as either a document or a community will be about that of a pile of shit.	Wikipedia talk pages	Cyberbullying (Aggression)
You are not worth the effort. You are arguing like Viriditas and Pename now. 24 hours really means 24 and a half hours. Four reverts in more than 24 hours is violating the "spirit" of the three revert in 24 hour rule - as interpreted by you. "So tough." Who needs rules? Just make it up as you go along and do what you want - call it "discretion". You violate the rules by blocking me then claim I violated the "spirit of the rule." Your violation is "debatable" like the Occupied Territories are "disputed." You are just abusing your authority to push your petty authoritarian agenda that obviously reflects your personal insecurities. You think you can threaten and bully me. "And I guess you won't be reverting so quickly in future, will you now¿' What a weasel. Please go ahead and contribute your petty complaints to ban me so I don't bother wasting my time on a project populated by immature arrogant twerps, fascist Zionist bigots, Islamophobe hate-mongers, bunch of lamea** bigots and losers. Why waste my time?	Wikipedia talk pages	Cyberbullying (Attack)

TABLE 5. Datasets used in the reviewed cyberbullying detection literature.

Source	Dataset	Total samples	Positive samples	Negative samples	Focus	Available	Papers
Twitter	Twitter-DS 1	12,705	391 (3%)	12314 (97%)	Cyberbullying		[12]
	Twitter-DS 2	16,014	5,355 (33%)	11,559 (72%)	Hate speech		[8]
	Twitter-DS 3	14,742	3,370 (23%)	11423 (77%)	Racism	✓	[81]
	Twitter-DS 4	1,762	685 (38.8%)	1,078 (61.18%)	Cyberbullying		[9]
	Twitter-DS 5	4,865	93 (3%)	4,700 (98%)	Cyberbullying		[82]
	Twitter-DS 6	296,308	-	-	Cyberbullying		[10], [11]
	Twitter-DS 7	7,321	2,102 (28.7%)	5,219 (71.2%)	Cyberbullying		[83], [84]
	Twitter-DS 8	9,484	-	-	Cyber-aggression	✓	[7]
	Twitter-DS 9	16,000	5,074 (31.6%)	10,926 (68.28%)	Cyberbullying		[21]
	Twitter-DS 10	14,194	1,753 (12.3%)	12,441 (87.7%)	Cyberbullying		[85]
	Twitter-DS 11	2,435	414 (17%)	2,021 (83%)	Hate speech	✓	[13]
	Twitter-DS 12	10,041	850 (10%)	9,191 (90%)	Cyberbullying	✓	[78]
	SemEval-DS	12722	5313 (42%)	7410 (58%)	Hate speech	✓	[86]
ASK.FM	Ask-DS	2,863,801	-	-	Cyberbullying		[10], [11], [87]
MySpace	Myspace-DS	3,245	950 (29.3%)	2,295 (60.7%)	Cyberbullying	✓	[63], [83]
Instagram	Instagram-DS 1	2,218 MS	665 (30%)	1,553 (70%)	Cyberbullying		[14], [16]
	Instagram-DS 2	9,828,760 MS	2,948,628 (30%)	6,880,132 (70%)	Cyberbullying	✓	[10], [11], [14]
	Instagram-DS 3	13,350 MS	1,602 (12%)	11,748 (88%)	Cyberbullying		[15]
	Instagram-DS 4	1,656,236 MS	-	-	Cyberbullying		[11]
Vine	Vine-DS 1	969 MS	303 (31%)	666 (69%)	Cyberbullying	✓	[47]
	Vine-DS 2	959 MS	45 (5%)	914 (95%)	Cyberbullying	✓	[16], [88]
FormSpring	FormSpring-DS 1	13,652	792 (6%)	12,860 (94%)	Cyberbullying	✓	[20]
	FormSpring-DS 2	12,000	825 (7%)	11,175 (93%)	Cyberbullying		[21]
	FormSpring-DS 3	13,160	2,205 (17%)	10,955 (83%)	Cyberbullying	✓	[53], [55]
YouTube	YouTube-DS 1	50,000 MS	-	-	Cyberbullying		[19]
	YouTube-DS 2	3,603 (users)	432 (12%)	3,171 (88%)	Cyberbullying	✓	[17]
	YouTube-DS 3	7,962	-	-	Cyberbullying		[18]
Wikipedia Talk Pages	Wikipedia-DS	115,737	13,542 (11.7%)	102,195 (88.3%)	Personal attacks	✓	[21], [66]
Reddit	Reddit-DS	10,100	-	-	Toxicity		[89]
Yahoo	Yahoo-Finance-DS	759,402	53,516 (7%)	705,886 (93%)	Abusive language		[77]
	Yahoo-News-DS	1,390,774	228,119 (16%)	1,162,655 (84%)	Abusive language		[77]

Note: A dash (-) denotes unavailable information. The "Available" column denotes whether the dataset is available for download either online or by contacting the authors.

The posts included in the dataset were crawled from MySpace's groups' feature and were manually labeled as normal or bullying-related.

- **YouTube** is an online video-sharing platform, which opens the door for cyberbullying as users can comment on the videos of other users. Keryov and Evelyn argue

that when YouTube videos are controversial, the comments tend to be more racist and abusive [91]. We found three studies that collected and used YouTube media sessions (videos + comments) to detect cyberbullying [17]–[19]. The dataset was generated by collecting media sessions on sensitive topics, like sexuality, race, culture, intelligence, and physical attributes.

- **Vine** was a short video hosting platform where users could share six-second long videos. Users were able to comment on those videos and sometimes the videos shared or the comments were racist towards certain groups of people. In 2018, Vine was archived and set to be replaced by a successor version but the project has been postponed indefinitely. Up until 2018, researchers could crawl data as media sessions (videos + comments) from Vine. Within the reviewed literature, we found two Vine datasets that were used in [47], [88] and [16].
- **Wikipedia Talk Pages** is a collaborative platform where Wikipedia users can discuss improvements on published articles on Wikipedia. Sometimes the comments are aggressive, toxic, and contain personal attacks. In the reviewed literature, we found one dataset that was collected by [66] and then used by [21]. Each comment in the dataset was labeled by 10 annotators via the Crowdflower (Figure-Eight) crowd-sourcing platform on whether it contains a personal attack.
- **Yahoo** is a web services provider that operates a number of different web services. We found only one study that used data from the Yahoo website [77]. They used comments posted on Yahoo Financial and Yahoo News stories for cyberbullying detection. All comments were moderated and annotated by Yahoo employees who were trained before the task in order to familiarise themselves with the required text judgment guidelines. In addition, the data are available for researchers.[10]
- **Reddit** is a popular social media network that offers social news aggregation, web content rating, and online discussions. [89] used comments posted on Reddit to detect triggers for toxicity. They focused on the ten subreddits with the highest number of subscribers. For each subreddit, they retrieved all the comments posted between January 2016 and August 2017 using Pushshift's public Reddit collection and used the Figure-Eight crowd-sourcing platform to label a subset of 10,100 randomly sampled comments from AskReddit.

In addition to the datasets in Table 5, Vidgen and Derczynski compiled a list of hate speech datasets [37]. That list[11] is a collection of annotated datasets for hate speech, online abuse, and offensive language. The collection contains datasets in different languages, e.g. Arabic, Croatian, Danish, English, French, German, Greek, Hindi-English, Indonesian, Italian, Polish, Portuguese, Slovene, Spanish and Turkish. The datasets were collected from different social media platforms, such as Twitter, Reddit, Facebook, Gab, and Wikipedia, and news platforms like Fox News and AlJazira.

From this list of the datasets used in the literature, we can see that Twitter is the most used platform for studying cyberbullying, which leads to a number of speculations, including that the moderation on Twitter is not so strict, it is an abundant source of bullying and hate or it is easier to retrieve data from because of the Twitter API. However, we believe that the cyberbullying detection community needs to release and use datasets that are collected from less mainstream platforms, but with even less strict moderation policies like an Urban Dictionary, 4&8 Chan, etc., because recent studies have shown that these platforms are often fertile ground for hate speech, and white supremacy [92], [93]. We also noticed that some of the platforms are now out of service, like Vine, or not any more popular, like ASK.FM, MySpace, and FormSpring. However, the data collected from these platforms are still relevant, as the offensive language is still the same and they can be used with more recent datasets to learn more about cyberbullying on social media.

2) DATA ANNOTATION

In the reviewed literature, we found two common ways the researchers used to label the collected data: i) manual annotation by humans, and ii) filtering using specific keywords. Manual annotation by humans is an arduous and time-consuming task. Some studies employed crowd-sourcing platforms to hire people without previous experience to label the data, in order to reduce the cost. Appen,[12] previously known as CrowdFlower, is one of the most used crowd-sourcing platforms and has been used by [7], [14], [21], [47], [66], [81]. Amazon Mechanical Turk (AMT)[13] is the second most used platform in the reviewed literature, used by [20], [21], [55]. Other studies hired experts to do the labeling. Some of those experts were linguists, e.g. [13], activist feminists [8], or experts in aggression in education systems [94]. Other studies hired graduate students to do the labelling [19], while in some other studies the researchers themselves did the labelling [18], [53], [77], [85].

To quantify the agreement between more than one annotator, researchers use the inter-annotators' agreement score, which can be measured using Cohen's kappa [95] or Krippendorff's alpha [96]. Crowd-sourcing platforms provide their agreement scores. The higher the score, the higher the agreement between annotators on whether the annotated item refers to cyberbullying or not. Among the studies that used crowd-sourcing platforms in the reviewed literature, the number of annotators hired to do the labelling was either three annotators [20], [21], [55], five annotators [7], [14], [47], [66], [81] or ten annotators [21], [66]. The inter-agreement scores, using Krippendorff's alpha or Cohen's kappa, between the annotators

[10] https://webscope.sandbox.yahoo.com/
[11] https://hatespeechdata.com/
[12] https://appen.com/
[13] https://www.mturk.com/

from the crowd-sourcing platforms ranged between 0.45 [7], [21], [66], 0.5 [14] and 0.79 [47], [88]. In the studies that hired experts to annotate the data, the number of hired experts ranged between one to two, given the increased cost compared to crowd-sourcing, with agreement scores reaching a Cohen's kappa of 0.78 [17] and a Cohen's kappa of 0.82 [85]. This indicates that despite the increased cost, experts are generally better at annotating the data. Nevertheless, crowd-sourced annotation can also provide high-quality results if the task is well designed to minimize confusion and eliminate unreliable annotators, eventually achieving reasonable agreement scores [47], [88].

When the filtering approach is used for labeling data, the available data are filtered using specific cyberbullying-related keywords and the matched data are labeled as referring to cyberbullying [9], [10], [15]. Filtering data using keywords could be unreliable, as some people may use profane words in a disguised or a friendly way, e.g. s**t [97]. In other cases, some people use high trending hashtags, which might be insulting words, to attract people to advertisement tweets.

As a result, even with keyword filtration, it is still useful to have a human annotator involved in labeling the data. However, an additional challenge exists. As often happens with subjective topics like cyberbullying, it is sometimes hard to tell if a post is an act of bullying or it is sarcastic. Consequently, more than one annotator is required, ideally an odd number, in order to reach a consensus in cases of disagreement.

3) DATASET SIZE AND BALANCE

Table 5 summarises all the datasets used in the reviewed literature and includes the size of the datasets and, whenever available, the number of positive samples (posts that include a form of cyberbullying) and the number of negative samples (posts that do not include any form of cyberbullying). One of the main challenges in automated cyberbullying detection is the availability of cyberbullying-related data. From the datasets in Table 5, we can see that seven datasets contain 10% or less of cyberbullying-related (positive) samples [12], [20], [21], [77], [78], [82], [88], while only one dataset is almost balanced, with 42% positive samples and 58% negative samples [86]. Nine datasets have a percentage of positive samples between 11.7% and 29% [13], [15], [17], [53], [66], [77], [83]–[85], while the rest of the datasets contain between 30% and 39% of positive samples [8], [9], [14], [47].

The imbalance in the datasets available in the literature may have a negative effect on using deep learning models. In the next section, we review some techniques used in the literature to address this imbalance in the datasets.

4) DATA SAMPLING

The imbalance of the datasets resulted in many researchers processing the datasets in order to ensure that the trained machine learning models learn to differentiate between cyberbullying cases and non-cyberbullying-related cases. Some works over-sampled the positive samples either by duplicating the positive samples multiple times in order to balance the dataset [20], [21], while other studies did the opposite by down-sampling negative samples in the dataset [53], [55], [82]. Some studies used search keywords on the streaming APIs to filter the incoming data and make sure to get more data with offensive content [12], [13], [66]. Others used Snowball sampling to ensure that they achieve a better representation of positive samples in the datasets [10], [88]. Krasnowska-Kieras *et al.* increased the number of positive samples by artificially generating cyberbullying-related tweets [78]. The rest of the studies opted to use the available imbalanced data to train their machine learning models, given that in a real-world situation, the number of cyberbullying-related posts is in general less than the number of other posts.

Even though over-sampling or under-sampling datasets could mitigate the imbalances in the datasets, they come with their own challenges. Because if not done properly, they could lead to over-fitting, as we will discuss in Section V. To mitigate these challenges, data augmentation could be used to generate more positive (bullying) text and balance the datasets.

In this section, we presented all the steps related to datasets in the cyberbullying detection literature. All these steps are important to ensure that the datasets are representative and less biased, in order to train fairer and generalizable models. In the next section, we review the next step in the machine learning pipeline, which is data pre-processing to clean the data and prepare them for training the ML model.

B. PRE-PROCESSING

Pre-processing is an important standard step for cleaning the data. In the reviewed literature, most of the works used the NLTK library[14] to tokenize, remove stop words, remove unwanted characters, correct misspelling, lemmatize and/or stem the raw data [98]–[102]. In the case of the Twitter datasets, more steps were typically applied, like replacing user mentions, URLs, and hashtags with special characters, as well as removing duplicates [22], [84], [103]. Some studies also used Part-of-Speech (POS) tagging as a pre-processing step [48], [98].

Even though these steps are almost identical in the literature, following these steps should depend on the task and the model used. For example, removing stop words is a standard step in most NLP applications, but in the case of cyberbullying detection, second and third nouns could be important indicators and features for cyberbullying, and removing them means losing important information (e.g., the word "f*ck" on its own is not necessarily used for bullying, contrary to being used in combination with a pronoun, such as "f*ck you"). Also more recent pre-trained models, like BERT, require a change in the pre-processing steps, as stemming is not needed anymore and punctuation symbols are important for

[14]https://www.nltk.org/

TABLE 6. Features used for automated cyberbullying detection in the reviewed literature and highest performance reported by each work.

Paper	Text Features	User Information Features	Sentiment Features	Word Embeddings	Other Features	Accuracy	Precision	Recall	F1	AUC
[19]	✓		✓			0.80				
[20]	✓	✓						0.87		
[17]	✓	✓	✓		✓	0.76				
[52]	✓				✓	0.88				
[12]	✓		✓			0.48				
[47]	✓		✓			0.76				
[8]	✓	✓					0.72	0.77	0.73	
[77]	✓			✓					0.81	
[81]	✓	✓		✓			0.92	0.92	0.91	
[81]	✓			✓	✓		0.76	0.79	0.78	
[82]	✓	✓							0.64	
[10]	✓				✓					0.83
[66]	✓								0.75	0.83
[83]	✓								0.68	0.80
[7]	✓	✓	✓					0.89	0.91	0.90
[21]	✓		✓		✓			0.92	0.91	0.91
[85]	✓									0.89
[13]	✓		✓			✓				0.92
[88]	✓	✓								0.68
[53]	✓									0.81
[11]	✓				✓			0.6		
[78]	✓				✓	✓				0.83
[15]	✓		✓				0.40	0.35	0.37	
[16]		✓	✓	✓					0.98	
[18]		✓	✓			✓	0.83	0.82	0.83	
[55]	✓						0.85	0.86	0.84	

the model to perform well, as shown in [104] where BERT is fine-tuned on tweets.

The next step in the pipeline after collecting, labeling, and pre-processing the data is the extraction of features that will be used for training the ML model.

C. FEATURES

In the reviewed literature, the most common features used can be grouped in the following four categories: 1) Text-based features, 2) User and Social media network information, 3) Sentiment and Psychological features, and 4) Distributional representation (word embeddings). We also considered one additional category called "Other features" to group some less common features used in some studies. A summary of the features used by different studies in the reviewed literature is provided in Table 6, while an overview of each feature category is provided below:

1) TEXT-BASED FEATURES

As shown in Table 6, text-based features are the most commonly used features in the reviewed literature. They are either used on their own or in combination with other features. Text features capture the patterns that exist in the text, which the machine learning models can then use to learn from the data. Various types of text features have been proposed in the literature, like the Bag of Words (BOW) models, which include one-hot encoding, Term Frequency (TF), and Term Frequency–Inverse Document Frequency (TF-IDF) representations. BOW with word N-grams is the most popular text representation model used in the reviewed literature [11], [13], [17]–[21], [47], [52], [53], [66], [77], [81], [83], [85], [88], [90]. Some studies used BOW with character N-grams and reported better results compared to the word N-grams BOW model [8], [21], [66], [77], [78], [81]. Other studies used the frequency of profane or negative words as features [12], [17]–[20], [88], while [12] used the frequency of the word "you" as a feature for detecting cyberbullying. Other studies used the number of words in the sentence (an online post), the number of hashtags used, the number of words in uppercase letters and the number of URLs in addition to the text [7], [12], [13], [15], [17], [81]. Furthermore, some studies applied natural language processing techniques and used Part-of-Speech (POS) tags related to the text as additional text features [19], [81]–[83].

2) USER INFORMATION

Besides using text-related information for feature selection, researchers have tried to use information related to the author of the examined text, and as a consequence, related to the person committing cyberbullying. This information could be the users' gender, age, or the number of their online posts, which can be found on their social media profiles.

In the reviewed literature, we found that gender has been used as a feature [8], [81], as according to [8], men tend to send more racist and sexist posts on Twitter than women. Anonymity is another factor that some researchers took into account since they claimed that users who are cyberbullies tend to hide their identities. However, results showed that it is not necessarily the case [17], [20]. Other researchers used information about the users' online behavior, like the number of their posts, their subscriptions, uploads, and their history of used words [17], [81], [88]. Furthermore, the users' location has also been used as a feature [8], [16]. User features also include information related to the user's social media network, like the users' number of followers, the number of likes and views they receive, or the number of people they follow [7], [16], [82], [88].

3) SENTIMENT AND PSYCHOLOGICAL FEATURES

Sentiment analysis refers to the task of using natural language processing and text analysis in order to evaluate the sentiment conveyed by a text, by assigning a sentiment score to the examined text. Positive scores typically relate to positive sentiment, while negative scores are typically indicative of negative sentiment [105]–[109]. In the reviewed literature, some researchers used the sentiment score of the text as a feature for cyberbullying detection, as negative words are an indicator of unpleasant and potentially bullying-related text [7], [12], [13], [15], [19], [21], [47], [83]. Some studies generated a sentiment score of the emotion icons (emojis) in the text and used those scores as features in training the machine learning models [7], [17].

In 2015, [110] developed a tool (LIWC[15]) that can analyze a text and reveal some of the psychological features of the author(s). For example, given that one of the main characteristics of a bully is to have power over their victims, the tool can be used to measure someone's tendency to exercise authority from their text. In the reviewed literature, [15] and [16] used the results of the LIWC tool as an additional feature for cyberbullying detection.

4) DISTRIBUTIONAL REPRESENTATION (WORD EMBEDDINGS)

A distributional text representation (Word embeddings) aims at representing words in a way that preserves their semantic relationships and takes into account the order of the words in the text [111]. Word embeddings have been widely used in recent years for most text classification and information retrieval tasks [112], [113]. However, there are few studies that used word embeddings in cyberbullying detection. Nobata *et al.* used the word2vec-CBOW word embedding to train their cyberbullying detection model [77]. Similarly, Agrawal *et al.* used Glove-Wikipedia to improve the task of cyberbullying detection [21]. Doc2vec embeddings were used as features in detecting cyberbullying by [11], who also adopted the idea of distributed representation of words and applied it to the user's online social network and developed node2vec as a feature for detecting cyberbullying. Koufakou *et al.* used FastText word embeddings that were retro-fitted for the task of cyberbullying detection [114], while other studies developed specialized word embeddings for the task of cyberbullying detection [9], [11], [55], [78].

In addition to the classic pre-trained models on Wikipedia and Google news, there have been new models pre-trained on Twitter, like Glove-Twitter,[16] Urban dictionary word embeddings pre-trained on words and definitions from the Urban Dictionary website [115], and Chan word embeddings pre-trained on text from the 4 & 8 Chan websites [116]. Despite these embeddings been trained with text that resembles more the way users communicate in social media platforms compared to the news and Wikipedia articles, the use of these embeddings has not yet been explored for the detection of cyberbullying.

5) OTHER FEATURES

Apart from the aforementioned features that were used in multiple studies, the following less common features were also used in the reviewed literature. Davdar *et al.* [17] hired experts to rate the importance of the extracted text features from the text and used this rating as an additional feature. They also used the information resulting from a multi-criteria decision support system (MCES) [117] as another feature to detect cyberbullying. Potha and Maragoudakis [52] used time series modeling and Singular Value Decomposition (SVD) to extract features for cyberbullying detection, and [9] used Latent Semantic Analysis (LSA), which is a topic modeling method, to extract different topics in the unlabelled text as features. Topic models like K-means, LDA, and LSI were also used in [118] to group the text into clusters and use this information as features. [16] used the metadata of images posted on social media platforms as features along with the time of the post. Few studies used Multi-modal cyberbullying detection, where the model is trained on both images and text to detect cyberbullying [76], [119]. [99] used the Levenshtein distance to measure the difference between two words as a feature to detect profane words in disguised form, e.g. f***, while [64] used the conditional feature probability to measure the importance of the features. [120] used a novel algorithm to reduce the number of features (text and user information) used in the classification task. They achieved an F1-score (will be discussed in Section IV-E) of 0.76 with an average of 6.6 features, compared to the baseline which achieved a 0.58 F1-score with 13 features.

6) FEATURE SELECTION

Singh *et al.* [82] proposed a method for combining text features, user features, and social media network features in a way that enhances the model's performance, by first determining the agreement score between different types of features and then determining the confidence score of certain

[15]http://liwc.wpengine.com

[16]https://nlp.stanford.edu/projects/glove/

feature types by calculating their accuracy in predicting the data label (as cyberbullying or not) from previous predictions. This way the model can determine which features are more important for each data instance and consequently make better predictions of the final data label. Using this approach, they achieved better results than other studies that combine features mindlessly, reporting an F1-score of 0.64.

Raisi and Huang used multi-view learning to maximize the mutual agreement across different features types (text and social network) of unlabelled data [11]. They used an ensemble of two learners: one to examine the language content of a post and another to consider the network structure of the post sender. They achieved a precision of 0.6, which is a relatively high score given that they do not use labeled data. Similarly, [16] built their model using multi-modality learning in order to use different pieces of information provided in the social media post, like images, videos, user profile, time, and location, assuming that the different pieces of information (modalities) could be complementary and achieved an F1-score of 0.98.

In the same direction of enhancing the learning of the different types of features, [83] proposed a framework called Sentiment Informed Cyberbullying Detection (SICD), which is a model that maximizes the use of sentiment information available in the post. They used the distribution of sentiment scores in the data to differentiate between the sentiment of cyberbullying posts and normal posts, achieving an AUC score of 0.80 and an F1-score of 0.68.

In this section, we reviewed the literature on the different features used in the task of cyberbullying detection. The most common features are Text-based and User information features. On the other hand, word embeddings are among the least used features even though they have been proven to perform well on several NLP tasks. The community of cyberbullying detection needs to explore more the use of word embeddings, especially with the release of the new contextual word embeddings like BERT and ELMO. We provide in-depth analysis and suggestions regarding feature selection for cyberbullying detection in Section V. We also implemented some of these suggestions in Section VI-D. In the next section, we reviewed the different ML models that have been used in the literature for the task of cyberbullying detection.

D. MACHINE LEARNING MODELS

In this section, we discuss the different ML models used for the task of cyberbullying detection in the reviewed literature.

1) RULES-BASED LEARNING

Some studies in the reviewed literature used rules-based models besides machine learning models to provide the criteria based on which the model classifies the data. They are especially used in the early studies, with less available training datasets than required to train machine learning models [12], [19], [20].

2) CONVENTIONAL MACHINE LEARNING

Conventional machine learning models are the most widely used in the reviewed literature. We found 61 (57.5%) studies that used conventional machine learning models. Most of them used supervised learning models. Among these supervised models are the models that are famous for performing well in text classification tasks, like Support Vector Machines (SVM) [9], [16], [19]–[21], [47], [52] and Naive Bayes (NB) [17], [19], [21], [47]. Other well-known models used are: Logistic Regression (LR) [8], [66], [88], Decision Trees (DT) [7], [17], [19], [20], [47], k-Nearest Neighbours (kNN) [18], [20], and Random Forests (RF) [7], [15], [16], [18], [21], [47]. Furthermore, despite the shortage of labelled datasets, there have been only a few trials that attempted to use weakly supervised [10], [11] or unsupervised machine learning models [53].

3) DEEP LEARNING

During the last two decades, deep learning models have been increasingly used in different variations and for different applications of machine learning. However, in the reviewed literature, we found that deep learning models have been used for cyberbullying detection much later. This could be because deep learning models need large numbers of data points for training and the available datasets for cyberbullying used to be small in numbers and in size, something that started to increase only recently. Zhao and Mao [100] used Semantic Enhanced Marginalised Denoising Auto-Encoder (smSDA) for cyberbullying detection. [21] used Transfer Learning with LSTM to detect cyberbullying across multiple social media platforms. They achieved a Precision score of 0.92, a Recall score of 0.91, and an F1-score of 0.91. CNN's have been also used to improve the detection of cyberbullying [13], [21], [22], [55], [68], [85], [122]. [21] and [11] used Long Short Term Memory (LSTM) models, which are a variation of Recurrent Neural Network (RNN) [23] models, to detect cyberbullying. [13] combined CNN layers with Gated Recurrent Network (GRN) layers to create a model for hate speech detection. Simpler deep learning models have also been explored in the literature. Some studies also used a simple neural network like the multi-layer perceptron (MLP) [66], [78].

4) UNCONVENTIONAL MODELS

Most of the reviewed papers used conventional machine learning models or deep learning models with a novel contribution in providing labeled datasets or in feature engineering. However, there are less common machine learning approaches like unsupervised learning, which have been used in other fields, e.g. for detecting spammer groups [123] and for rumor detection [124], [125], or semi-supervised machine learning models [126], [127]. These unconventional methods have also been used for cyberbullying detection. [128] and [88] proposed a multi-stage cyberbullying detection model that improves the classification time by 223 times

over the baseline and the time needed to raise an alert is improved seven times over the baseline, achieving a precision of 0.71 and a recall of 0.66. [83] first used a distant supervised based sentiment machine learning model to measure the sentiment score distribution of the dataset and then they incorporated that score to detect cyberbullying. They reported an AUC score of 0.80 and an F1-score of 0.68. [53] used Fuzzy Finger Prints to identify the unique fingerprints of the positive cyberbullying examples in the training dataset. They slightly outperformed the baselines for unbalanced datasets and achieved an F1-score of 0.77. [49] used hierarchical attention networks to mirror the structure of social media sessions and use attention mechanisms that capture the relationship between the words in a comment within a certain context, achieving an F1-score of 0.78 and an AUC score of 0.851.

In this section, we reviewed the different models used in the literature on cyberbullying detection. We can see that the majority of the studies, reviewed here, opted for conventional ML models over deep learning models which could be due to the small sizes of the datasets and the high imbalance ratio of positive (bullying) and negative (not-bullying) data. The more datasets being released for the task of cyberbullying detection, the more deep learning models will be easier to use. From our experiments, described in Section VI, RNN models, especially Bidirectional LSTM models, look like the most promising deep learning models for the task of cyberbullying detection. We also noticed that the literature is missing out on new advances in pre-trained language models like BERT, GPT2, and GPT3. In the next section, we review the different evaluation methods used in the literature of cyberbullying and their validity.

E. EVALUATION METRICS

Given the use of machine learning for cyberbullying detection in the reviewed literature, the performance of the reviewed methods was evaluated using typical evaluation metrics that are common across the machine learning literature. The majority of the examined works used the following evaluation metrics: accuracy, F1-score, precision, recall (also known as sensitivity or true positive rate), as well as Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) scores. These metrics are computed based on the four outcomes that summarise a binary classification task's results, i.e. i) True Positive (TP), the number of correctly classified positive samples, ii) True Negative (TN), the number of correctly classified negative samples, iii) False Positive (FP), the number of samples miss-classified as positive, and iv) False Negative (FN), the number of samples miss-classified as negative. In addition, a few works reported the error score or the Mean Squared Error (MSE) score [12], [52].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Despite being one of the most common metrics for classification, accuracy (Eq. 1) is not the preferred evaluation metric when working with imbalanced datasets [129], since it may lead to overestimated scores as a result of a high number of samples belonging to a certain class. In the reviewed cyberbullying detection literature, we found that [19], [47] and [18] used the accuracy metric to report the results of their models, while studies that used deep learning did not report accuracy scores.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

Some of the studies that reported precision (Eq. 2), also reported recall (Eq. 3) and F1-score (Eq. 4) [8], [9], [18], [21], [55]. Other works reported only the F1-score [13], [15], [16], [53], [82], [83], [88], [90], or either recall only [20] or precision and recall [94].

AUC is generally preferred in binary classification tasks but despite cyberbullying detection being a binary classification task, we found few studies in the reviewed literature that reported AUC scores, either on their own or along the F1-score [10], [17], [66], [78], [83], [85]. A summary of the reviewed studies that reported an AUC score or an F1-score higher than 0.80 is provided in Table 7, including the achieved scores, the dataset, the features, and the machine learning models used.

In this section, we reviewed the different evaluation metrics used in the literature of cyberbullying detection. We showed that using accuracy is not advisable for tasks where there is a high imbalance in the dataset. We also recommend the use of the F1-score as a good measure of the models' ability to find a balance between precision and recall.

In the next section, we provide an analysis of the limitations in the literature of cyberbullying detection and provide some recommendations to overcome these limitations.

V. LIMITATIONS OF THE REVIEWED LITERATURE

Examining the reviewed literature, it is evident that there are some limitations and challenges in the field of cyberbullying detection in terms of the datasets, features, machine learning models, and evaluation approaches used.

A. DATASET-RELATED CHALLENGES

Some of the challenges that make the task of cyberbullying detection harder are related to the cyberbullying datasets available in the literature and are mostly related to the definition of cyberbullying, to data annotation, class imbalance, underlying biases, and language. In this section, we discuss these challenges.

1) DEFINITION

The lack of a clear distinction in the definition between cyberbullying and related concepts, like hate speech, affects the generalisability of the state-of-the-art models proposed in

TABLE 7. The best F1 and AUC scores achieved in the reviewed literature. The evaluation scores presented here are for providing an idea of the scores being reported in the literature but are not meant for comparative reasons as these studies used different datasets.

Paper	Dataset (size)	Features	Model	AUC	F1
[69]	Kaggle-insults (4000)	Text Features Word Embeddings	Support Vector Machine (SVM)	0.85	-
[59]	Twitter (1900)	Text Features User Features Other Features	Sequential Minimal Optimisation (SMO)	0.96	-
[10]	Twitter (296,308)	Text Features	Participant Vocabulary Consistency (PVC)	0.83	-
[7]	Twitter (9,484)	Text Features User Features Network Features Sentiment Features Word Embeddings	Random Forest (RF)	0.90	-
[21]	Twitter (16,000)	Text Features Sentiment Features Word Embeddings	Long Short Term Memory (LSTM)	0.93	-
[105]	MySpace (600)	Text Features Other Features	Naive Bayes (NB)	-	0.89
[121]	MySpace (-)	Text Features	NB + Use Fuzzy rule based + Genatic algorithm	-	0.98
[50]	Twitter (10007)	Text Features User Features Network Features Psycological Features Other Features	SVM	-	0.94
[22]	Formspringme (13000)	Text Features Sentiment Features Word Embeddings Other Features	Convolution Neural Network (CNN)	0.98	0.98
[85]	Visr child safety data (-)	Text Features Psychological Features	CNN	0.89	-
[13]	Twitter (2,435)	Text Features Sentiment Features	CNN	0.92	-
[55]	FormSpring (13160)	Word Embeddings	CNN	-	0.84
[78]	Twitter (10,041)	Text Features Word Embeddings	NN	0.83	-
[16]	Instagram (155,267)	Psychological Features User Features Network Features Image meta data Time	RF	-	0.98
[66]	Wikipedia Talk Pages (115,737)	Text Features	Logistic Regression (LR) Multi Layer Perception (MLP)	0.96*	-

Note: * refers to ROC-AUC

the literature. It also affects the choice of features that can be used to enhance the models' performance in detecting cyberbullying or hate speech. For example, Fortuna *et al.* [34] suggest that there are two types of features, general textual-based features and specific hate speech based features. Some of these features intersect with cyberbullying detection like *Othering Language* and *Perpetrator Characteristics (e.g. gender and geographic localisation)*, while others are specific for the task of hate speech detection, like *Declaration of superiority of the group, Focus on particular stereotypes, and Intersectionism of oppression.* The lack of a clear definition of the detection task makes it harder to select the most suitable features and models from the literature.

2) ANNOTATIONS

We found that the studies that used crowd-sourcing platforms to annotate the data reported low inter-agreement scores among the annotators. This could be due to a lack of clear instructions given to the annotators or due to the demographic of the annotators which may lead to unknown biases [79]. These biases and low agreement scores may cause over-fitting in the models reported in the literature, which in turn affects their generalisability. Related information about the annotators' demographics was not shared or described in the reviewed papers. To address this issue, we recommend that future studies share this information along with the data description when a new dataset is released.

3) CLASS IMBALANCE

The statistics presented in Figure 3 show a clear pattern of imbalance between the number of positive (abusive) data samples and the number of negative (normal) data samples in the datasets used in the reviewed literature. This imbalance imposes some limitations on the use of deep learning models. To overcome this problem, some studies over-sample the positive samples in the dataset, which, if done before the

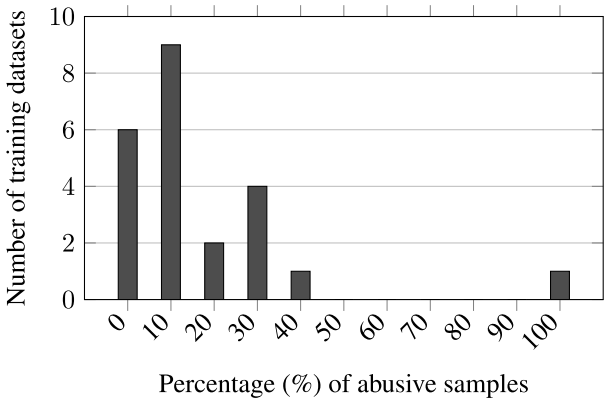


FIGURE 3. Histogram of the percentage of abusive samples in the reviewed datasets in the literature.

train-test split, becomes problematic and causes model overfitting, as demonstrated by [79].

4) USER DISTRIBUTION BIAS

There is, potentially, a user distribution bias in the datasets used in the literature. For example, one of the most used datasets in the literature of cyberbullying and hate speech detection is the tweets dataset collected by Wassem *et al.* [8]. The dataset contains 14K tweets annotated as ''racist'', ''sexist'' or ''none''. The number of hateful tweets (sexist and racist) is 4,839 and the number of non-hateful tweets is 10,110. The data on the users who generated these tweets were analyzed by [79], who found that all the data was generated by 1,590 users, with 491 users having generated all the sexist tweets and only 8 users have generated all the racist tweets. Among the ''sexist'' tweets, 40% were generated by a single user and among the ''racist'' tweets, 90% were generated by a single user. Furthermore, they argued that the models trained on Wassem *et al.*'s dataset are prone to overfitting due to the user distribution.

5) LANGUAGE

Despite languages other than English having been included in the datasets found in the literature, these ''language'' datasets are limited in sources to almost only Twitter and Facebook. There is a clear lack of ''language'' datasets that cover other social media platforms. Furthermore, most of these ''language'' datasets contain hate speech, and very few contain cyberbullying and its sub-types. As a consequence, this limits the research on cyberbullying detection in languages other than English. There is a need for more cyberbullying datasets in other languages to advance the research and improve the detection of cyberbullying in these languages.

B. FEATURES-RELATED CHALLENGES

Identified challenges in relation to the features used for cyberbullying detection are related to the lack of use of visual features, the word embeddings used for text representation, and the availability of user and network information.

1) VISUAL FEATURES

Table 6 summarises the most common features used in the literature to detect cyberbullying. From this table, it is evident that the use of visual features for cyberbullying detection is rare [14], [130], [131]. As recent studies have shown that teenagers make extensive use of visual content on platforms like Instagram and Snapchat for their communication [132], [133], it is important to develop models that can detect cyberbullying from visual media, in order to provide a form of protection to the receivers of such visual content.

2) TEXT REPRESENTATION

Another limitation we found in the literature is the use of relevant word embeddings to the task of cyberbullying detection. As discussed earlier, the main word embeddings used in the literature are Word2Vec, Glove, or Doc2Vec. However, more recent word embeddings have been proposed that may be more relevant to the task, such as sentiment specific word embedding (SSWE) [134] and Urban Dictionary word embedding [135]. Aragwal and Awekar experimented with different deep learning models trained with different word embeddings like Glove and SSWE and found that the performance of the models trained with Glove and SSWE is very close [21]. However, they did not conduct any intrinsic analysis to compare the semantic relatedness of SSWE and Glove to cyberbullying datasets. Similarly, contextual word embeddings like ELMO, GPT, and BERT [136]–[138] have not been explored enough in the literature. We recommend using the new advances in NLP to improve the detection of cyberbullying.

3) USER AND NETWORK INFORMATION

Although some studies in the reviewed literature used user and network information as features to detect cyberbullying, few studies share this user information, which is limiting to the development of the field. This may be partially attributed to the general data protection regulations. However, for an important task such as cyberbullying detection, it would be more beneficial to share this information in an anonymized form than not providing it at all.

C. MACHINE LEARNING MODELS-RELATED CHALLENGES

After reviewing the literature on the machine learning models used to detect cyberbullying and their training process, we identified challenges related to the generalisability of the models and the lack of use of new advances in NLP, like attention-based models and transfer learning.

1) MODEL GENERALISABILITY

The first challenge is the validity of the results reported in the literature, as Arango *et al.* showed in their study on the generalisability of prior work on the detection of hate speech and cyberbullying [79]. They showed through a series of experiments that the models that are used as state-of-the-art in the literature of cyberbullying detection failed to generalize

to new datasets, which means that the high scores reported in the original papers are due to over-fitting. They explain that the over-fitting occurs due to some mistakes in the training process: 1) Extracting the features from the whole dataset (training and test sets) for training instead of extracting the features only from the training set; 2) Oversampling the positive (abusive) content to balance the dataset before the train-test split; 3) Bias resulting from the uneven distribution of the users who generate the abusive content within the dataset. Their findings suggest that we should look at the results reported in the literature with a critical view and carefully assess the reported training processes. We also recommend replicating the results of the models reported in the literature before using them.

2) CONTEXTUAL LANGUAGE MODELS

The second challenge is that although attention-based mechanisms and pre-trained models like ELMO, GPT, and BERT [136]–[138] have been around for quite some time now, there are few studies that used these models to detect cyberbullying or hate speech [139]–[141]. Pre-trained models like BERT have established a new state of the art in many NLP tasks, requiring only small datasets to fine-tune the model on the downstream tasks [142].

3) TRANSFER LEARNING

Transfer learning is a great technique to mitigate the issue of small and imbalanced datasets which, as discussed earlier, is a problem with the task of cyberbullying detection. It can also be beneficial in training a model that can detect different types of cyberbullying regardless of the data source. However, transfer learning has not been widely explored in the community of cyberbullying detection except for few studies [143]–[145].

D. EVALUATION-RELATED CHALLENGES

1) OVER-FITTING

As mentioned earlier, some studies report high F1-scores like, e.g., 0.934 and 0.961 [21], [146]. However, [79] showed that these high F1-scores are due to over-fitting, as discussed in Section V-C1. To address this issue, we recommend testing any model's generalisability and report the performance on an unseen dataset besides reporting the performance results on the test set. For example, the SemEval 2019 [86] dataset could be used for that reason if the task is hate speech detection. However, we acknowledge that the lack of cyberbullying datasets can be an obstacle to achieving that.

2) METRICS

We found some studies in the reviewed literature that reported classification accuracy for assessing performance, which is not reliable when working with unbalanced databases, such as the ones typically available in the cyberbullying and hate speech literature. Considering the very high proportion of negative (non-abusive) samples in the available datasets, the high accuracy values are biased towards the high number of true-negatives in the test set. For NLP tasks it is best to report the F1-score in order to get a realistic evaluation of a model's performance [147], [148].

VI. IMPROVING THE PERFORMANCE OF CYBERBULLYING DETECTION

After reviewing the literature on automated cyberbullying detection and identifying its major challenges and limitations, in this section we attempt to address some of these challenges. We first replicated one of the studies that reported high performance. Then, we extended the experiment to train and test the proposed method from the replicated study on other cyberbullying-related datasets from different data sources. Afterward, we aimed to address some of the limitations discussed in Section V by using BERT and slang-based word embeddings to improve the detection of cyberbullying.

A. STUDY REPLICATION

We opted to replicate the study by Wulczyn *et al.* [66], as they report the highest ROC AUC scores among the reviewed literature and their dataset is available online. In that study, the authors used machine learning models to detect cyberbullying from the comments sent on the Wikipedia Talk Pages (WTP) platform. They used the Appen crowd-sourcing platform to label the training dataset, by hiring ten workers on Appen to answer the following set of questions for each comment:

1) Does this comment contain a personal attack or harassment?
2) How friendly or aggressive is this comment?
3) Rate the toxicity of this comment.

The inter-rater agreement (Krippendorff's alpha) score between the workers was 0.45. The dataset contains 115,737 comments with 13,542 (11.7%) of the comments containing forms of personal attacks.

Then, they used a conventional machine learning model, Logistic Regression (LR), and a deep learning model, Multi-Layer Perceptron (MLP), with word N-grams of (1,2) features and characters with N-grams of (1,5) features. They also used two types of class labels. The first is One-Hot (OH) binary labels, which means that if the comment is considered aggression by the majority of the crowd workers then the label is 1, and if the majority of the crowd workers consider the comment, not aggression, then the label is 0. The second type of label is Empirical Distribution (ED), which takes into consideration the distribution of answers of the crowd workers on the comments. For example, if a comment was labeled as aggressive by seven crowd workers and labeled as non-aggressive by three workers, then the label of this comment would be represented as (0.7, 0.3). Finally, they used the ROC AUC score to evaluate the performance of the models only on the personal attacks dataset, which is the

TABLE 8. Results reported in the Wulczyn *et al.* [66] study.

				ROC AUC	
Dataset	Samples	Positive samples	N-gram type	LR	MLP
WTP (Attack)	115,864	13,590 (11.7%)	Word	0.95	0.95
			Char	0.92	0.95

TABLE 9. Results for the replicated Wulczyn *et al.* [66] study.

		ROC AUC		AUC		F1-score	
Dataset	N-gram type	LR	MLP	LR	MLP	LR	MLP
WTP (Attack)	Word	0.95	0.95	0.82	0.78	0.74	0.70
	Char	0.92	0.95	0.758	0.79	0.64	0.69

dataset that contains answers for the question "*Does this comment contain personal attack or harassment*?". The results acquired from the original study are reported in Table 8, reaching the highest ROC AUC of 0.95.

In our replication of the study, we only used One-Hot labels because only One-Hot labels are available for the other examined datasets. Furthermore, in addition to ROC AUC scores, we also reported the AUC and F1 scores. In addition, the Scikit-learn[17] and Keras [149] Python packages were used for the implementation of the replication study. The results of the replication study are reported in Table 9 and support the results reported in the original paper (Table 8). The reported results show that MLP, either with char N-grams or word N-grams, provides the best ROC AUC score of 0.95, as also reported in the original study. However, when we used the integer predictions instead of prediction probabilities to measure the AUC, we find that LR with the word N-grams model is the best performing, achieving an AUC score of 0.82. Furthermore, the LR with the word N-grams model also achieved the highest F1-score of 0.74. In the next sections, we investigated the performance of the replicated study when trained and tested on different cyberbullying datasets from different sources and with different types of cyberbullying.

B. DETECTING OTHER TYPES OF CYBERBULLYING

To evaluate the ability of the replicated study to detect other types of cyberbullying, we trained and tested the replicated model on three additional datasets from different sources:

- Twitter: Two collections of Twitter messages collected by [8] for hate speech detection, who defines hate speech as targeting individuals or groups on the basis of their characteristics, demonstrating a clear intention to incite harm, or to promote hatred, and may or may not use offensive or profane words. The first collection is *Twitter-Racism*, which contains racist comments, while the second collection is *Twitter-Sexism*, which contains sexist comments.
- Kaggle-insults: The *Kaggle-insults* dataset is part of a Kaggle competition[18] and contains insults described as: "Insults could contain profanity, racial slurs, or other offensive languages. But oftentimes, they do not".

Before training the examined machine learning models, the datasets were first pre-processed using the NLTK package [150] to tokenize the text and convert all words into lower case. Then, the Porter Stemming algorithm [151] was used to stem the words, and then stop words, punctuation, and numbers were removed. Finally, weblinks and some words that are frequently found in tweets like "amp" and "lol" were also removed. After pre-processing, each dataset was randomly split into a training set (70%) and test set (30%) for training and evaluating the models proposed in the replicated study using the new datasets. Results are reported in Table 10 in terms of the F1-score.

TABLE 10. F1-scores of the replicated study on the additional datasets.

Dataset	Size	Positive samples	N-gram type	LR	MLP
Kaggle (Insults)	7557	2649 (35%)	Word	**0.67**	0
			Char	0.65	0
Twitter (Racism)	13471	1970 (14.5%)	Word	**0.74**	0
			Char	0.70	0
Twitter (Sexism)	14881	3377 (22.6%)	Word	**0.72**	0
			Char	0.69	0

Note: Bold values indicate best performance per dataset. For MLP models, all samples were predicted as belonging to the negative class, leading to an F1-score equal to 0

From Table 9 and 10, it is evident that the performance of the replicated LR model with word N-grams on its original dataset (WTP Personal attack) (F1-score = 0.74) is close to its performance on the Twitter-racism dataset (F1-score = 0.74) and the Twitter-Sexism dataset (F1-score = 0.72), while the performance drops for the Kaggle-insults dataset, achieving an F1-score of 0.67. The MLP models underperformed considerably on all of the three examined datasets. Interestingly, in all cases the MLP models predicted all samples as belonging to the majority negative (no cyberbullying) class, resulting in an F1-score equal to 0. We speculate that this low performance is due to the size of the datasets which are significantly smaller than the WTP-Attack dataset that has 115,864 samples compared to the 14,881 samples of the Twitter-Sexism dataset, which is the biggest of the three examined datasets, as shown in Table 10. The performance of the MLP model was significantly improved after adding a trainable embedding layer, as explained in the following section and shown in Table 11.

C. CONTEXTUAL LANGUAGE MODELS - BERT

As discussed earlier, pre-trained contextual language models are not widely used yet for the task of cyberbullying detection. In this section, we tested the performance of fine-tuned BERT

[17] https://scikit-learn.org/stable/

[18] https://www.kaggle.com/c/detecting-insults-in-social-commentary/data

TABLE 11. F1-scores for LR and updated MLP versus state-of-the-art deep learning models and BERT on the three examined datasets.

Dataset	LR	MLP	LSTM	Bi-LSTM	BERT
Kaggle-insults	0.670	0.572	0.642	0.653	**0.768**
Twitter-racism	0.740	0.688	0.640	0.678	**0.747**
Twitter-sexism	0.720	0.669	0.6569	0.649	**0.760**

Note: Bold values indicate best performance per dataset.

on the three examined datasets and compared its performance to the replicated study and other state-of-the-art deep learning models. To fine-tune BERT on the examined datasets, we first applied different pre-processing steps than the ones described in Section VI-B in order to make the most out of BERT's pre-training. We followed the pre-processing steps used in [104] where BERT was fine-tuned on tweets: 1) We removed URLs, user mentions, and non-ASCII characters. For Twitter datasets, we also removed the retweet abbreviation ''RT''. 2) All letters were lowercased. 3) Contractions were converted to their formal format. 4) Space was added between words and punctuation marks.

We fine-tuned BERT for the task of text classification on the examined datasets by employing the BERT_{base}(uncased) [138] model. For fine-tuning, the BERT model was trained for 10 epochs with a batch size of 32 and a learning rate of $2e^{-5}$, as suggested in [138]. The sequence length parameter changed across datasets depending on the maximum token length of each dataset. For the Twitter-sexism and Twitter-racism datasets, a sequence length of 64 was used (maximum observed sequence length in the dataset), while 128 (the maximum we could use due to available computational resources limitations) was used for the Kaggle-insults dataset. A single linear layer was added on top of the pooled output of BERT for the final text classification.

For Logistic Regression and MLP, we used the same models as in the replicated study but added a trainable embedding layer to the MLP model to improve performance. It must be noted that the word N-grams was used in both cases since they provided the best performance in both the original and the replicated experiments. We also used two state-of-the-art deep learning models, LSTM [152] and Bidirectional LSTM [153], with the same architecture as in [21], who used RNN models to detect cyberbullying, employing the same pre-processing steps like the ones described in Section VI-B. To this end, we first used the Keras tokenizer [149] to convert the text into numerical vectors (each integer is the index of a token in a dictionary) with a maximum length of 600 (the maximum we could use due to available computational resources limitations) for the Kaggle-insults dataset and 41 (maximum observed sequence length in the dataset) for the Twitter datasets. Similar to the MLP model, a trainable embedding layer was used as the first layer with an input size equal to a given dataset's vocabulary size and an output size equal to the size of the embedding. Then, the embedding layer was fed to the Bi-LSTM model. To avoid over-fitting, we used L2 regularisation with a value of 10^{-7} which gave the best results after experimenting with different values. The two models were then trained for 100 epochs with a batch size of 32, using the Adam optimizer and a learning rate equal to 0.01, which is the default of the Keras optimizer.

Results in terms of the classification F1-scores are provided in Table 11 for all the examined models and datasets. It is evident that BERT significantly outperformed the models from the replicated study (LR model, updated MLP model), as well as the state-of-the-art LSTM and Bi-LSTM models on all three tested datasets, achieving the highest performance on the Kaggle-insults dataset with an F1-score of 0.768 and the lowest performance on Twitter-racism with an F1-score of 0.747. The Friedman statistical significance test [154] was then used to statistically compare the models' performance, showing that BERT significantly outperformed all the models ($p < 0.05$). These results demonstrate that BERT significantly improves performance on the task of cyberbullying detection compared to other widely used methods for text classification. Interestingly, Linear Regression provided the second-best performance, outperforming the updated MLP, LSTM, and Bi-LSTM models.

D. SLANG-BASED WORD EMBEDDINGS

As we discussed earlier, the available literature on cyberbullying detection has not exploited some of the recent pre-trained word embeddings that have been shown to increase performance in various NLP tasks. These word embeddings are not pre-trained on the news or Wikipedia articles but are instead pre-trained on social media data, like text collected from Twitter or the Urban Dictionary. We call these word embeddings ''slang-based word embeddings''. We hypothesize that because these word embeddings are trained on a text that resembles more the way users communicate online compared to how news or Wikipedia articles are written, they may improve the detection of cyberbullying compared to other word embeddings that were typically pre-trained on the news or Wikipedia articles.

To evaluate our hypothesis, we used recently released word embeddings that were pre-trained on slang-based datasets like the Urban Dictionary (UD), Sentiment Specific Word Embedding (SSWE), Glove-Twitter (Glv-Twtr), and Glove-Common Crawl (Glv-CC) and compared them to random weight initialization (RI), traditional word embeddings like Word2vector (W2V) pre-trained on news articles, and Glove embeddings pre-trained on Wikipedia articles (Glv-WK). These embeddings were used with the MLP, LSTM, and Bi-LSTM models with similar settings as the ones described in Section VI-C, in order to evaluate whether the use of slang-based embeddings would lead to increased performance. Results in terms of the F1-score are provided in Table 12.

As shown in Table 12, the results for the Kaggle-insults dataset show that for MLP and Bi-LSTM, the slang-based Urban Dictionary (UD) word-embedding performed the best with F1-scores of 0.692 and 0.694 respectively. The performance of the Bi-LSTM model using the Glove-Twitter (Glv-Twtr) slang-based word embedding is also the same as

TABLE 12. F1-scores for the MLP, LSTM, and Bi-LSTM models using the examined word embeddings for the three datasets. Bold indicates the best performance per model and dataset.

Dataset	Model	Word Embeddings						
		RI	Glv_WK	Glv_CC	Glv_Twtr	SSWE	UD	W2V
Kaggle-insults	MLP	0.654	0.587	0.655	0.672	0.534	**0.692**	0.654
	LSTM	0.654	0.697	**0.710**	0.699	0.563	0.672	0.686
	Bi-LSTM	0.680	0.690	0.655	**0.694**	0.627	**0.694**	0.691
Twitter-Racism	MLP	0.688	0.688	0.693	**0.696**	0.595	0.679	0.638
	LSTM	0.661	0.680	**0.696**	0.693	0.671	0.673	0.662
	Bi-LSTM	0.655	0.691	0.685	**0.692**	0.680	0.683	0.669
Twitter-Sexism	MLP	0.664	0.625	0.669	0.665	**0.673**	0.657	0.617
	LSTM	0.686	0.676	0.696	0.680	0.672	**0.696**	0.678
	Bi-LSTM	0.665	0.694	**0.701**	0.678	0.630	0.692	0.669

when the UD word embeddings are used (F1-score = 0.694). For the LSTM model, the slang-based Glove-Common Crawl (Glv-CC) word embedding provided the best performance with an F1-score of 0.710.

For the Twitter-Racism dataset, the slang-based Glove-Twitter (Glv-Twtr) word embedding provided the best performance for the MLP and Bi-LSTM models, achieving F1-scores of 0.696 and 0.692 respectively. For the LSTM model, the best performing word embedding was the slang-based Glove-common Crawl (Glv-CC), resulting in an F1-score of 0.696.

Finally, for the Twitter-Sexism dataset, the slang-based SSWE word embedding provided the best performance for the MLP model (F1-score = 0.673), the slang based word embedding UD provided the best performance for the LSTM model (F1-score = 0.696), and the slang-based Glove-Common Crawl (Glv-CC) word embeddings provided the best performance for the Bi-LSTM model (F1-score = 0.701).

From Table 12, it is evident that in all cases, slang-based word embeddings provided the best performance. We used the Friedman statistical test to compare the F1-scores of all the seven word embeddings (RI, Glv-WK, Glv-CC, Glv-Twtr, SSWE, UD, and W2V) for each model across all the datasets, but a statistically significant difference could not be established ($p > 0.05$). Nevertheless, the consistently better performance of slang-based word embeddings in all the examined cases shows that using the slang-based word embeddings, especially Glove-Common Crawl, Glove-Twitter, Urban Dictionary, and Sentiment Specific Word Embedding (SSWE), can improve the performance of cyberbullying detection. Nevertheless, despite the enhanced performance of the MLP, LSTM, and Bi-LSTM models when combined with slang-based word embeddings, the fine-tuned BERT model achieved the best F1 scores across all the examined datasets. An interesting future work direction would be to pre-train BERT on cyberbullying-related data or slang-based data, as it could be very beneficial and lead to potential performance improvements on the task of cyberbullying detection.

VII. CONCLUSION

In this work, we conducted a systematic literature review on automated cyberbullying detection. The motivation behind this area of research is to help in preventing cyberbullying and its negative consequences that can include depression, low self-esteem and even committing suicide. We organized the reviewed literature around the steps of the machine learning pipeline employed by each reviewed work, due to the lack of a similar systematic study in the literature. In the reviewed literature, we identified some challenges and limitations of the available work on cyberbullying detection, some of which are related to the cyberbullying datasets used in the various works. In particular, challenges with defining cyberbullying, the annotation of datasets, data imbalance, data bias, and limited availability of multi-lingual datasets. We also noticed that the literature is not up-to-date with using more recent slang-based word embeddings like the urban dictionary word embeddings; with using more recent models; with using contextual language models like BERT; and with using transfer learning. Another limitation relates to the use of classification accuracy as a performance evaluation metric which can be deceiving when there is an imbalance in the datasets.

In the second part of this work, we conducted a series of experiments to address some of the identified limitations and investigate their impact on the task of cyberbullying detection. Our results demonstrate that using contextual-based language models like BERT significantly improved the detection of cyberbullying in comparison to the chosen replicated study and state-of-the-art deep learning models. We also found that deep learning models with a trainable embedding layer initialized with slang-based pre-trained word embeddings outperformed random initialization and traditional pre-trained word embeddings like Word2Vec pre-trained on news articles and Glove trained on Wikipedia articles. Nevertheless, the fine-tuned BERT model still outperformed all the examined deep learning models even with the slang-based word embeddings, demonstrating its potential for the task of cyberbullying detection.

For future work, we are considering exploring the pre-training of BERT on slang-based text in addition to Wikipedia

articles and the Books Corpus which may improve its performance on the task of cyberbullying detection even more.

REFERENCES

[1] S. Z. Omar, A. Daud, M. S. Hassan, J. Bolong, and M. Teimmouri, "Children internet usage: Opportunities for self development," *Procedia Social Behav. Sci.*, vol. 155, pp. 75–80, Nov. 2014.

[2] L. Haddon and S. Livingstone, "Risks, opportunities, and risky opportunities: How children make sense of the online environment," in *Cognitive Development in Digital Contexts*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 275–302.

[3] T. K. H. Chan, C. M. K. Cheung, and R. Y. M. Wong, "Cyberbullying on social networking sites: The crime opportunity and affordance perspectives," *J. Manage. Inf. Syst.*, vol. 36, no. 2, pp. 574–609, Apr. 2019.

[4] M. Duggan, "Online harassment 2017," Pew Res. Center, Washington, DC, USA, Tech. Rep., Jul. 2017.

[5] G. M. Abaido, "Cyberbullying on social media platforms among university students in the united arab emirates," *Int. J. Adolescence Youth*, vol. 25, no. 1, pp. 407–420, Dec. 2020, doi: 10.1080/02673843.2019.1669059.

[6] F. Sticca, S. Ruggieri, F. Alsaker, and S. Perren, "Longitudinal risk factors for cyberbullying in adolescence," *J. Community Appl. Social Psychol.*, vol. 23, no. 1, pp. 52–67, Jan. 2013.

[7] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," in *Proc. ACM Conf. Web Sci. (WebSci)*. New York, NY, USA: Association for Computing Machinery, 2017, pp. 13–22.

[8] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93, doi: 10.18653/v1/n16-2013.

[9] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," ser. ICDCN '16. New York, NY, USA: Association for Computing Machinery, 2016, doi: 10.1145/2833312.2849567.

[10] E. Raisi and B. Huang, "Cyberbullying detection with weakly supervised machine learning," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2017, pp. 409–416.

[11] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 479–486.

[12] L. P. D. Bosque and S. E. G. Villareal, "Aggressive text detection for cyberbullying," in *Proc. 13th Mex. Int. Conf. Artif. Intell. Hum.-Inspired Comput. Appl. (MICAI)* in Lecture Notes in Computer Science, vol. 8856, Nov. 2014, A. F. Gelbukh, F. Castro-Espinoza, and S. N. Galicia-Haro, Eds. Tuxtla Gutiérrez, Mexico: Springer, 2014, pp. 221–232, doi: 10.1007/978-3-319-13647-9_21.

[13] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-GRU based deep neural network," in *The Semantic Web*, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds. Cham, Switzerland: Springer, 2018, pp. 745–760.

[14] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in *Proc. 7th Int. Conf., (SocInfo)* in Lecture Notes in Computer Science, vol. 9471, T. Liu, C. N. Scollon, and W. Zhu, Eds. Beijing, China: Springer, Dec. 2015, pp. 49–66, doi: 10.1007/978-3-319-27433-1_4.

[15] H.-T. Kao, S. Yan, D. Huang, N. Bartley, H. Hosseinmardi, and E. Ferrara, "Understanding cyberbullying on instagram and Ask.Fm via social role detection," in *Proc. Companion World Wide Web Conf.*, May 2019, pp. 183–188, doi: 10.1145/3308560.3316505.

[16] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in *Proc. 12th ACM Int. Conf. Web Search Data Mining (WSDM)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 339–347, doi: 10.1145/3289600.3291037.

[17] M. Dadvar, D. Trieschnigg, and F. de Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies," in *Proc. 27th Can. Conf. Artif. Intell. Adv. Artif. Intell. Can. (AI)* in Lecture Notes in Computer Science, vol. 8436, M. Sokolova and P. van Beek, Eds. Montréal, QC, Canada: Springer, May 2014, pp. 275–281, doi: 10.1007/978-3-319-06483-3_25.

[18] A. Kumar, S. Nayak, and N. Chandra, "Empirical analysis of supervised machine learning techniques for cyberbullying detection," in *Proc. Int. Conf. Innov. Comput. Commun.*, S. Bhattacharyya, A. E. Hassanien, D. Gupta, A. Khanna, and I. Pan, Eds. Singapore: Springer, 2019, pp. 223–230.

[19] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. Social Mobile Web, Papers (ICWSM) Workshop*, Barcelona, Spain, Jul. 2011, pp. 1–7. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3841

[20] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA)*, vol. 2, Dec. 2011, pp. 241–244, doi: 10.1109/ICMLA.2011.152.

[21] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *Advances in Information Retrieval*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham, Switzerland: Springer, 2018, pp. 141–153.

[22] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon, "Cyberbullying detection with a pronunciation based convolutional neural network," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 740–745.

[23] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, vol. 2, 2010, pp. 1045–1048.

[24] A. Nocentini, V. Zambuto, and E. Menesini, "Anti-bullying programs and information and communication technologies (ICTs): A systematic review," *Aggression Violent Behav.*, vol. 23, pp. 52–60, Jul./Aug. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359178915000749

[25] N. M. Zainudin, K. H. Zainal, N. A. Hasbullah, N. A. Wahab, and S. Ramli, "A review on cyberbullying in malaysia from digital forensic perspective," in *Proc. Int. Conf. Inf. Commun. Technol. (ICICTM)*, 2016, pp. 246–250.

[26] B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying detection: A survey on multilingual techniques," in *Proc. Eur. Model. Symp. (EMS)*, Nov. 2016, pp. 165–171.

[27] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 3–24, Jan. 2020.

[28] N. Tarwani, U. Chorasia, and P. K. Shukla, "Survey of cyberbulling detection on social media big-data," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, pp. 831–835, 2017.

[29] T. Mahlangu, C. Tu, and P. Owolawi, "A review of automated detection methods for cyberbullying," in *Proc. Int. Conf. Intell. Innov. Comput. Appl. (ICONIC)*, Dec. 2018, pp. 1–5.

[30] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. Veiga Simão, and I. Trancoso, "Automatic cyberbullying detection: A systematic review," *Comput. Hum. Behav.*, vol. 93, pp. 333–345, Apr. 2019.

[31] N. Tahmasbi and A. Fuchsberger, "Challenges and future directions of automated cyberbullying detection," Amer. Conf. Inf. Syst., USA, Tech. Rep. 9780996683166, 2018.

[32] M. Dadvar and K. Eckert, *Cyberbullying Detection in Social Networks Using Deep Learning Based Models; a Reproducibility Study*. New York, NY, USA: Springer, Dec. 2018, doi: 10.13140/RG.2.2.16187.87846.

[33] S. Nadali, M. A. A. Murad, N. M. Sharef, A. Mustapha, and S. Shojaee, "A review of cyberbullying detection: An overview," in *Proc. 13th Int. Conf. Intellient Syst. Design Appl.*, Dec. 2013, pp. 325–330.

[34] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, Sep. 2018.

[35] M. A. Al-garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," *IEEE Access*, vol. 7, pp. 70701–70718, 2019.

[36] C. Emmery, B. Verhoeven, G. De Pauw, G. Jacobs, C. Van Hee, E. Lefever, B. Desmet, V. Hoste, and W. Daelemans, "Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity," 2019, *arXiv:1910.11922*. [Online]. Available: http://arxiv.org/abs/1910.11922

[37] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," *PLoS ONE*, vol. 15, no. 12, Dec. 2020, Art. no. e0243300.

[38] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: A systematic review," *Lang. Resour. Eval.*, vol. 55, no. 2, pp. 477–523, Jun. 2021, doi: 10.1007/s10579-020-09502-8.

[39] M. Mladenović, V. Ošmjanski, and S. V. Stanković, "Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–42, Apr. 2021, doi: 10.1145/3424246.

[40] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Inf. Sci.*, vol. 497, pp. 38–55, Sep. 2019.

[41] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[42] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[43] S. Raschka, "Naive Bayes and text classification i–introduction and theory," 2014, *arXiv:1410.5329*. [Online]. Available: http://arxiv.org/abs/1410.5329

[44] J. W. Patchin and S. Hinduja, *Cyberbullying Prevention and Response: Expert Perspectives*. Evanston, IL, USA: Routledge, 2012.

[45] R. M. Kowalski, S. P. Limber, and P. W. Agatston, *Cyberbullying: Bullying in the Digital Age*. Hoboken, NJ, USA: Wiley, 2012.

[46] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *J. Child Psychol. Psychiatry*, vol. 49, no. 4, pp. 376–385, Apr. 2008.

[47] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in vine," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2015, pp. 617–622, doi: 10.1145/2808797.2809381.

[48] C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," *PLoS ONE*, vol. 13, no. 10, Oct. 2018, Art. no. e0203794.

[49] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the instagram social network," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2019, pp. 235–243.

[50] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433–443, Oct. 2016, doi: 10.1016/j.chb.2016.05.051.

[51] B. Belsey, "Cyberbullying: An emerging threat to the 'always on' generation," *Recuperado el*, vol. 5, no. 5, p. 2010, 2005.

[52] N. Potha and M. Maragoudakis, "Cyberbullying detection using time series modeling," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Dec. 2014, pp. 373–382, doi: 10.1109/ICDMW.2014.170.

[53] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur, "Using fuzzy fingerprints for cyberbullying detection in social networks," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2018, pp. 1–7.

[54] S. Parime and V. Suri, "Cyberbullying detection and prevention: Data mining and psychological perspective," in *Proc. Int. Conf. Circuits, Power Comput. Technol. [ICCPCT]*, Mar. 2014, pp. 1541–1547.

[55] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, "A 'deeper' look at detecting cyberbullying in social networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[56] M. Munezero, M. Mozgovoy, T. Kakkonen, V. Klyuev, and E. Sutinen, "Antisocial behavior corpus for harmful language detection," in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, Sep. 2013, pp. 261–265.

[57] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and H. H. Reese, "Cyber bullying among college students: Evidence from multiple domains of college life," in *Misbehavior Online in Higher Education*. Bingley, U.K.: Emerald Group Publishing Limited, 2012.

[58] T. Bosse and S. Stam, "A normative agent system to prevent cyberbullying," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Aug. 2011, pp. 425–430.

[59] P. Galán-García, J. G. De La Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying," *Logic J. IGPL*. vol. 24, no. 1, pp. 42–53, 2015.

[60] R. Pawar, Y. Agrawal, A. Joshi, R. Gorrepati, and R. R. Raje, "Cyberbullying detection system with multiple server configurations," in *Proc. IEEE Int. Conf. Electro/Inf. Technol. (EIT)*, May 2018, pp. 0090–0095.

[61] S. Hinduja and J. W. Patchin, "Cyberbullying: Identification," in *Prevention and Response, Cyberbullying Research Center*. London, U.K.: Routledge, 2014.

[62] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: Query terms and techniques," in *Proc. 5th Annu. ACM Web Sci. Conf. (WebSci)*, 2013, pp. 195–204.

[63] J. Bayzick, "Detecting the presence of cyberbullying using computer software," Honors thesis, Dept. Math. Comput. Sci., Ursinus College, Collegeville, PA, USA, 2011.

[64] I. Nazar, D.-S. Zois, and M. Yao, "A hierarchical approach for timely cyberbullying detection," in *Proc. IEEE Data Sci. Workshop (DSW)*, Jun. 2019, pp. 190–195.

[65] M. Duggan, L. Rainie, A. Smith, C. Funk, A. Lenhart, and M. Madden, "Online harassment. Washington, DC: Pew research center," Pew Res. Center, Washington, DC, USA, Tech. Rep., Oct. 2014.

[66] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1391–1399.

[67] R. Shetgiri, "Bullying and victimization among children," *Adv. Pediatrics*, vol. 60, no. 1, p. 33, 2013.

[68] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of cyberbullying using deep neural network," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2019, pp. 604–607.

[69] V. S. Chavan and S. S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Aug. 2015, pp. 2354–2358.

[70] W. Romsaiyud, K. na Nakornphanom, P. Prasertsilp, P. Nurarak, and P. Konglerd, "Automated cyberbullying detection using clustering appearance patterns," in *Proc. 9th Int. Conf. Knowl. Smart Technol. (KST)*, Feb. 2017, pp. 242–247.

[71] C. Chelmis, D.-S. Zois, and M. Yao, "Mining patterns of cyberbullying on Twitter," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 126–133.

[72] L. Cheng, J. Li, Y. Silva, D. Hall, and H. Liu, "PI-bully: Personalized cyberbullying detection with peer influence," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 5829–5835.

[73] P. K. Smith, "Cyberbullying and cyber aggression," in *Handbook of School Violence and School Safety*. Evanston, IL, USA: Routledge, 2012, pp. 111–121.

[74] X. Tian, "Investigating cyberbullying in social media: The case of Twitter," in *Proc. KSU Conf. Cybersecur. Educ., Res. Pract.*, Atlanta, GA, USA, 2016.

[75] A. Mahmud, K. Z. Ahmed, and M. Khan, "Detecting flames and insults in text," in *Proc. Int. Conf. Natural Lang. Process.* Dhaka, Bangladesh: BRAC Univ., 2008, pp. 1–11.

[76] K. B. Kansara and N. M. Shekokar, "A framework for cyberbullying detection in social network," *Int. J. Current Eng. Technol.*, vol. 5, no. 1, pp. 494–498, 2015.

[77] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 145–153, doi: 10.1145/2872427.2883062.

[78] K. Krasnowska-Kieras and A. Wróblewska, "A simple neural network for cyberbullying detection," in *Proc. PolEval Workshop*, 2019, pp. 161–163.

[79] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in *Proc. 42nd Int. ACM sigir Conf. Res. Develop. Inf. Retr.*, 2019, pp. 45–54.

[80] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 10, no. 1, pp. 1–4, 2016.

[81] Z. Waseem, "Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter," in *Proc. 1st Workshop NLP Comput. Social Sci.*, 2016, pp. 138–142. [Online]. Available: https://www.aclweb.org/anthology/W16-5618

[82] V. K. Singh, Q. Huang, and P. K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 884–887.

[83] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Machine Learning and Knowledge Discovery in Databases*, M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, Eds. Cham, Switzerland: Springer, 2017, pp. 52–67.

[84] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in *Proc. Conf. North Amer. chapter Assoc. Comput. linguistics, Hum. Lang. Technol.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 656–666.

[85] Q. Huang, D. Inkpen, J. Zhang, and D. Van Bruwaene, "Cyberbullying intervention based on convolutional neural networks," in *Proc. 1st Workshop Trolling, Aggression Cyberbullying (TRAC)*. Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 42–51. [Online]. Available: https://aclanthology.org/W18-4405

[86] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 54–63, doi: 10.18653/v1/s19-2007.

[87] H. H. S. Li, Z. Yang, Q. Lv, R. I. R. R. Han, and S. Mishra, "A comparison of common users across instagram and Ask.Fm to better understand cyberbullying," in *Proc. IEEE 4th Int. Conf. Big Data Cloud Comput.*, Dec. 2014, pp. 355–362, doi: 10.1109/BDCloud.2014.87.

[88] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra, "Scalable and timely detection of cyberbullying in online social networks," in *Proc. 33rd Annu. ACM Symp. Appl. Comput.*, Apr. 2018, pp. 1738–1747, doi: 10.1145/3167132.3167317.

[89] H. Almerekhi, H. Kwak, B. J. Jansen, and J. Salminen, "Detecting toxicity triggers in online discussions," in *Proc. 30th ACM Conf. Hypertext Social Media*, Sep. 2019, pp. 291–292.

[90] V. Nahar, X. Li, C. Pang, and Y. Zhang, "Cyberbullying detection based on text-stream classification," in *Proc. Conferences Res. Pract. Inf. Technol. Ser., Austral. Comput. Soc.*, vol. 146, 2013, pp. 49–58.

[91] E. Keryova, "YouTube: Online video and participatory culture," Wiley, U.K., Tech. Rep. 978-0-745-66019-6, 2020.

[92] A. Papasavva, S. Zannettou, E. De Cristofaro, G. Stringhini, and J. Blackburn, "Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board," 2020, *arXiv:2001.07487*. [Online]. Available: http://arxiv.org/abs/2001.07487

[93] D. Nguyen, B. McGillivray, and T. Yasseri, "Emo, love, and god: Making sense of urban dictionary, a crowd-sourced online dictionary," 2017, *arXiv:1712.08647*. [Online]. Available: http://arxiv.org/abs/1712.08647

[94] M. Ptaszynski, F. Masui, T. Nitta, S. Hatakeyama, Y. Kimura, R. Rzepka, and K. Araki, "Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization," *Int. J. Child-Comput. Interact.*, vol. 8, pp. 15–30, May 2016.

[95] L. Burla, B. Knierim, J. Barth, K. Liewald, M. Duetz, and T. Abel, "From text to codings: Intercoder reliability assessment in qualitative content analysis," *Nursing Res.*, vol. 57, no. 2, pp. 113–117, 2008.

[96] K. Krippendorff, "Computing krippendorff's alpha-reliability," Penn Libraries, Univ. Pennsylvania, Philadelphia, PA, USA, Tech. Rep., Jan. 2011.

[97] A. Tommasel, J. M. Rodriguez, and D. L. Godoy, "Features for detecting aggression in social media: An exploratory study," in *Proc. 19th Simposio Argentino de Inteligencia Artif. (ASAI)-JAIIO (CABA)*, 2018, pp. 1–14.

[98] U. Bretschneider, T. Wöhner, and R. Peters, "Detecting online harassment in social networks," in *Proc. 35th Int. Conf. Inf. Syst. (ICIS)*. Atlanta, GA, USA: Association for Information Systems, 2014, pp. 1–14.

[99] B. S. Nandhini and J. I. Sheeba, "Cyberbullying detection and classification using information retrieval algorithm," in *Proc. Int. Conf. Adv. Res. Comput. Sci. Eng. Technol. (ICARCSET)*, 2015, p. 20.

[100] R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 328–339, Jul. 2017.

[101] Y. J. Foong and M. Oussalah, "Cyberbullying system detection and analysis," in *Proc. Eur. Intell. Secur. Informat. Conf. (EISIC)*, Sep. 2017, pp. 40–46.

[102] H. K. Sharma, K. Kshitiz, and Shailendra, "NLP and machine learning techniques for detecting insulting comments on social networking platforms," in *Proc. Int. Conf. Adv. Comput. Commun. Eng. (ICACCE)*, Jun. 2018, pp. 265–272.

[103] S. Tomkins, L. Getoor, Y. Chen, and Y. Zhang, "A socio-linguistic model for cyberbullying detection," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 53–60.

[104] H. Dang, K. Lee, S. Henry, and O. Uzuner, "Ensemble bert for classifying medication-mentioning tweets," in *Proc. 5th Social Media Mining Health Appl. Workshop Shared Task*, 2020, pp. 37–41.

[105] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[106] R. D. Desai, "Sentiment analysis of Twitter data," in *Proc. 2nd Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, Jun. 2018, pp. 30–38.

[107] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process. (HLT)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 347–354.

[108] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing*, vol. 2. London, U.K.: Chapman & Hall, 2010, pp. 627–666.

[109] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. 7th Int. Conf. Lang. Resour. Eval. (LREC)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010, pp. 1–7. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

[110] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Univ. Texas, Austin, TX, USA, Tech. Rep., 2015.

[111] T. Mikolov, W. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proc. (NAACL-HLT)*, Nov. 2013, pp. 746–751. [Online]. Available: https://www.aclweb.org/anthology/N13-1090

[112] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. R. Kingsbury, and H. Liu, "A comparison of word embeddings for the biomedical natural language processing," *J. Biomed. Informat.*, vol. 87, pp. 12–20, Nov. 2018, doi: 10.1016/j.jbi.2018.09.008.

[113] C. Wang, P. Nulty, and D. Lillis, "A comparative study on word embeddings in deep learning for text classification," in *Proc. 4th Int. Conf. Natural Lang. Process. Inf. Retr.*, Dec. 2020, pp. 37–46, doi: 10.1145/3443279.3443304.

[114] A. Koufakou, V. Basile, and V. Patti, "Florunito@trac-2: Retrofitting word embeddings on an abusive lexicon for aggressive language detection," in *Proc. 2nd Workshop Trolling, Aggression Cyberbullying, (TRAC@LREC)*, May 2020, R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, and D. Kadar, Eds. Marseille, France: European Language Resources Association (ELRA), 2020, pp. 106–112. [Online]. Available: https://www.aclweb.org/anthology/2020.trac-1.17/

[115] S. R. Wilson, W. Magdy, B. McGillivray, K. Garimella, and G. Tyson, "Urban dictionary embeddings for slang NLP applications," in *Proc. 12th Lang. Resour. Eval. Conf., (LREC)*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4764–4773. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.586/

[116] P. Voué, T. De Smedt, and G. De Pauw, "4chan & 8chan embeddings," 2020, *arXiv:2005.06946*. [Online]. Available: http://arxiv.org/abs/2005.06946

[117] S. Zahir, "Making public policy decisions using a web-based multi-criteria electoral system (MCES)," *Int. J. Inf. Technol. Decis. Making*, vol. 1, no. 2, pp. 293–309, Jun. 2002, doi: 10.1142/S0219622002000178.

[118] E. Sutinen, "Automatic detection of antisocial behaviour in texts," *Informatica (Ljubljana)*, vol. 38, no. 1, pp. 1–8, 2014.

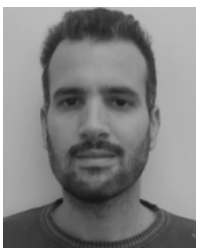
[119] G. NaliniPriya and M. Asswini, "A dynamic cognitive system for automatic detection and prevention of cyber-bullying attacks," *ARPN J. Eng. Appl. Sci.*, vol. 10, no. 10, pp. 4618–4626, 2015.

[120] M. Yao, C. Chelmis, and D.-S. Zois, "Cyberbullying detection on instagram with optimal online feature selection," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 401–408.

[121] B. S. Nandhini and J. I. Sheeba, "Online social network bullying detection using intelligence techniques," *Procedia Comput. Sci.*, vol. 45, pp. 485–492, Jan. 2015.

[122] M. A. Al-Ajlan and M. Ykhlef, "Optimized Twitter cyberbullying detection based on deep learning," in *Proc. 21st Saudi Comput. Soc. Nat. Comput. Conf. (NCC)*, Apr. 2018, pp. 1–5.
[123] S.-J. Ji, Q. Zhang, J. Li, D. K. W. Chiu, S. Xu, L. Yi, and M. Gong, "A burst-based unsupervised method for detecting review spammer groups," *Inf. Sci.*, vol. 536, pp. 454–469, Oct. 2020.
[124] S. M. Alzanin and A. M. Azmi, "Rumor detection in arabic tweets using semi-supervised and unsupervised expectation–maximization," *Knowl.-Based Syst.*, vol. 185, Dec. 2019, Art. no. 104945.
[125] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, "Unsupervised rumor detection based on users' behaviors using neural networks," *Pattern Recognit. Lett.*, vol. 105, pp. 226–233, Apr. 2018.
[126] X. Gu, "A self-training hierarchical prototype-based approach for semi-supervised classification," *Inf. Sci.*, vol. 535, pp. 204–224, Oct. 2020.
[127] R. A. R. Ashfaq, X.-Z. Wang, J. Z. Huang, H. Abbas, and Y.-L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system," *Inf. Sci.*, vol. 378, pp. 484–497, Feb. 2017.
[128] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra, "Investigating factors influencing the latency of cyberbullying detection," 2016, *arXiv:1611.05419*. [Online]. Available: http://arxiv.org/abs/1611.05419
[129] S. Rogers and M. Girolami, *A First Course in Machine Learning*, 2nd ed. London, U.K.: Chapman & Hall, 2016.
[130] V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, May 2017, pp. 2090–2099.
[131] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–26, 2018.
[132] J. A. Pater, A. D. Miller, and E. D. Mynatt, "This digital life: A neighborhood-based study of adolescents' lives online," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst.*, Apr. 2015, pp. 2305–2314.
[133] V. K. Singh, M. L. Radford, Q. Huang, and S. Furrer, "'They basically like destroyed the school one day' on newer app features and cyberbullying in schools," in *Proc. ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, 2017, pp. 1210–1216.
[134] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for Twitter sentiment classification," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1. Baltimore, MD, USA: Association for Computational Linguistics, Jun. 2014, pp. 1555–1565. [Online]. Available: https://www.aclweb.org/anthology/P14-1146
[135] S. Wilson, W. Magdy, B. McGillivray, K. Garimella, and G. Tyson, "Urban dictionary embeddings for slang NLP applications," in *Proc. 12th Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, May 2020, pp. 4764–4773. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.586
[136] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, *arXiv:1802.05365*. [Online]. Available: http://arxiv.org/abs/1802.05365
[137] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. (2018). *Improving Language Understanding by Generative Pre-Training*. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf
[138] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805
[139] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," *PLoS ONE*, vol. 14, no. 8, Aug. 2019, Art. no. e0221152.
[140] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained BERT model," in *Proc. Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, Jul. 2020, pp. 1096–1100.
[141] S. Paul and S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimedia Syst.*, vol. 710, pp. 1–8, Nov. 2020.
[142] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification," in *Proc. China Nat. Conf. Chin. Comput. Linguistics*. Kunming, China: Springer, 2019, pp. 194–206.
[143] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII*, H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, Eds. Cham, Switzerland: Springer, 2020, pp. 928–940.
[144] Z. Mossie, "Social media dark side content detection using transfer learning emphasis on hate and conflict," in *Proc. Companion Proc. Web Conf.*, Apr. 2020, pp. 259–263, doi: 10.1145/3366424.3382084.
[145] Z. Waseem, J. Thorne, and J. Bingel, *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*. Cham, Switzerland: Springer, 2018, pp. 29–55, doi: 10.1007/978-3-319-78583-7_3.
[146] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW) Companion*, 2017, pp. 759–760.
[147] S. Rogers and M. Girolami, *A First Course in Machine Learning*. Boca Raton, FL, USA: CRC Press, 2016.
[148] T. Joachims, "A statistical learning learning model of text classification for support vector machines," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2001, pp. 128–136.
[149] Tensorflow.org. (2020). *Text Tokenization Utility Class*. Accessed: Sep. 28, 2020. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/%Tokenizer
[150] NLTK Project. *Natural Language Toolkit*. Accessed: Sep. 28, 2020. [Online]. Available: https://www.nltk.org/
[151] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
[152] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[153] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 997.
[154] D. W. Zimmerman and B. D. Zumbo, "Relative power of the Wilcoxon test, the friedman test, and repeated-measures ANOVA on ranks," *J. Experim. Educ.*, vol. 62, no. 1, pp. 75–86, Jul. 1993.

FATMA ELSAFOURY received the B.Sc. degree (Hons.) in computing science from Helwan University, Egypt, in 2008, and the M.Sc. degree in data science from the University of Glasgow, U.K., in 2019. She is currently pursuing the Ph.D. degree in computing science with the University of the West of Scotland, U.K. She is also a Knowledge Transfer Partner Research Associate working on cyberbullying detection from collaborative platforms that are used by school students in Scotland. She has publications of social computing and participated in peer-reviewed for journals and conferences. Her research interests include studying the risks and opportunities of social media using machine learning and natural language processing tools.



STAMOS KATSIGIANNIS (Member, IEEE) received the B.Sc. degree (Hons.) in informatics and telecommunications from the National and Kapodistrian University of Athens, Greece, in 2009, the M.Sc. degree in computer science from the Athens University of Economics and Business, Greece, in 2011, and the Ph.D. degree in computer science (biomedical image and general purpose video processing) from the National and Kapodistrian University of Athens, in 2016. He is currently an Assistant Professor with the Department of Computer Science, Durham University, U.K. He has participated in six national and international research projects, and has authored and coauthored over 45 research publications, including peer-reviewed journals, book chapters, and conference proceedings. His research interests include affective computing, image analysis, machine learning, video coding, image and video quality, and GPU computing.

ZEESHAN PERVEZ (Senior Member, IEEE) is currently a Professor of computer science with the University of the West of Scotland (UWS). He has published over 80 indexed journals, peer-reviewed conferences, and book chapters. He is actively involved in various U.K./EU and international funded projects. He has a track record of securing substantial funding and delivering projects funded through H2020, Erasmus+, Innovate U.K., Knowledge Transfer Partnership (KTP), Microsoft Asia, Microsoft Research, Scottish Funding Council, Ministry of Knowledge Economy (South Korea), to name a few. His research interests include large-scale data analysis, data stream processing, the Internet of Things (IoT), cybersecurity, and cloud computing. He is a fellow of the Higher Education Academy, U.K. He is a full member of the EPSRC Peer Review College, U.K.

NAEEM RAMZAN (Senior Member, IEEE) received the M.Sc. degree in telecommunications from the University of Brest, France, in 2004, and the Ph.D. degree in electronics engineering from the Queen Mary University of London, London, U.K., in 2008. He is currently a Full Professor of artificial intelligence and the Director of the Affective and Human Computing for Smart Environment (AHCSE) Research Centre, University of the West of Scotland (UWS), U.K. He has authored or coauthored more than 200 research publications, including journals, book chapters, and standardization contributions. He has authored a book and co-edited some books as well. His research interests are cross-disciplinary and industry focused and include AI/machine learning, affective computing and multimedia processing, analysis and communication, video quality evaluation, brain-inspired multi-modal cognitive technology, big data analytics, affective computing, the IoT/smart environments, natural multi-modal human–computer interaction, and eHealth/connected Health. His article was awarded the Best Paper Award 2017 of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and number of conference papers were selected for the Best Student Paper Award. He has been a lead researcher in various nationally or EU sponsored multimillion funded international research projects (total funding as PI secured over £20m). He is a Senior Fellow of the Higher Education Academy (HEA), the Co-Chair of MPEG HEVC verification (AHG5) Group, and a Voting Member of the British Standard Institution (BSI). He has been awarded the Scottish Knowledge Exchange Champion Award 2020 and numerous other awards, such as the Staff Appreciation and Recognition Scheme (STARS) Award for Leadership, in 2019, the STARS Award 2015 and 2017 for Outstanding Research and Knowledge Exchange (the University of the West of Scotland), and the Contribution Reward Scheme 2011 and 2009 for outstanding research and teaching activities (the Queen Mary University of London). In addition, he holds key roles in the Video Quality Expert Group (VQEG), such as the Co-Chair of the Ultra High Definition (UltraHD) Group, the Co-Chair of the Visually Lossless Quality Analysis (VLQA) Group, and the Co-Chair of the Psycho-Physiological Quality Assessment (PsyPhyQA). He is also the Co-Editor-in-Chief of *VQEG eLetter*. He has served as a guest editor for a number of journals. He is also a Founding Associate Editor of *Journal of Quality and User Experience* (Springer) and an associate editor of number of journals. He has chaired/co-chaired/organized more than 25 workshops, special sessions, and tracks in international conferences. He has developed a highly innovative portfolio of post graduate studies, including the M.Sc. degrees in advanced computing, big data, the IoT, and eHealth/digital health.