International Journal of Information Security (2022) 21:1409–1431
https://doi.org/10.1007/s10207-022-00600-y

SURVEY

# A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media

Raju Kumar[1] · Aruna Bhat[1]

Accepted: 8 July 2022 / Published online: 8 August 2022

**Abstract**
Online social media (OSM) is an integral part of human life these days. Significantly, the young generation spends most of their time on social media in an active and passive state. The exponential growth of OSM has created an atmosphere of increased cybercrime. Although OSM provides a platform to connect people with similar thoughts and interests, it also exposes vulnerable users to mischievous elements in cyberspace. Social media connects and generates a massive amount of human activity-related data. However, the misuse of OSM introduces a novel way of expressing aggression and violence that exclusively happens online. In this research paper, we briefly discuss the background of Cyberbullying and the various machine and deep learning-based models incorporated to deal with it effectively. We also highlight the main challenges in designing a cyberbullying prediction model and address them.

**Keywords** Cyberbullying · Machine learning · Online harassment · Cyber forensics · Online social media

## 1 Introduction

OSM is gaining popularity day by day. Its user base ranges from the younger generation to the older generation worldwide. It not only helps us to connect but also provides a medium for sharing our thoughts and interests. However, some malicious users may use it for mischievous purposes like threatening, stalking, cyberbullying, harassing, etc. The victims of these activities sometimes get affected to such an extent that it. Unfortunately, it directly impacts their physical and mental health. They may face psychological issues like depression, anxiety, aggression, etc. In severe cases, it may also lead to suicidal tendencies. Due to this, many users feel unsafe and are afraid to utilize OSM's benefits. Therefore, it is evident that if the misuse of OSM is not controlled in a timely and effective manner, it may eventually lead to severe social problems [1].

Cyberbullying is a severe social problem where offenders harm others mentally and psychologically and have damaging effects. It creates a category of violence and aggression that happens only via online platforms, mainly social media [2]. Lately, cyberbullying has become an important area of research. Although most of the present study deals with cyberbullying effects and the victim's psychological state after the attacks, there has not been significant work in predicting these cybercrimes to control the bullying on social media. Recently devised techniques attempt to halt a cyberbullying attack or even prevent it during or before it happens. Many machine learning (ML)- and deep learning(DL)-based predictive models have been made to predict a potential bully and an aggressive post on social media.

OSM is not only used to connect people but also generates a massive amount of human behavior-related data, which may be textual or multimodal. These data work as fuel for predictive models. Based on a user's past data, we can predict their behavior on social media using various techniques. This survey paper mainly attempts to test and perform a comparative analysis of some of the significant existing ML- and DL-based predictive algorithms for cyberbullying and highlight the challenges faced by the researchers in designing. It was observed that ML, DL, and natural language processing (NLP) are dominant tools to counter bullying attacks [3] effectively.

✉ Aruna Bhat
aruna.bhat@dtu.ac.in

Raju Kumar
yadavraju03@yahoo.com

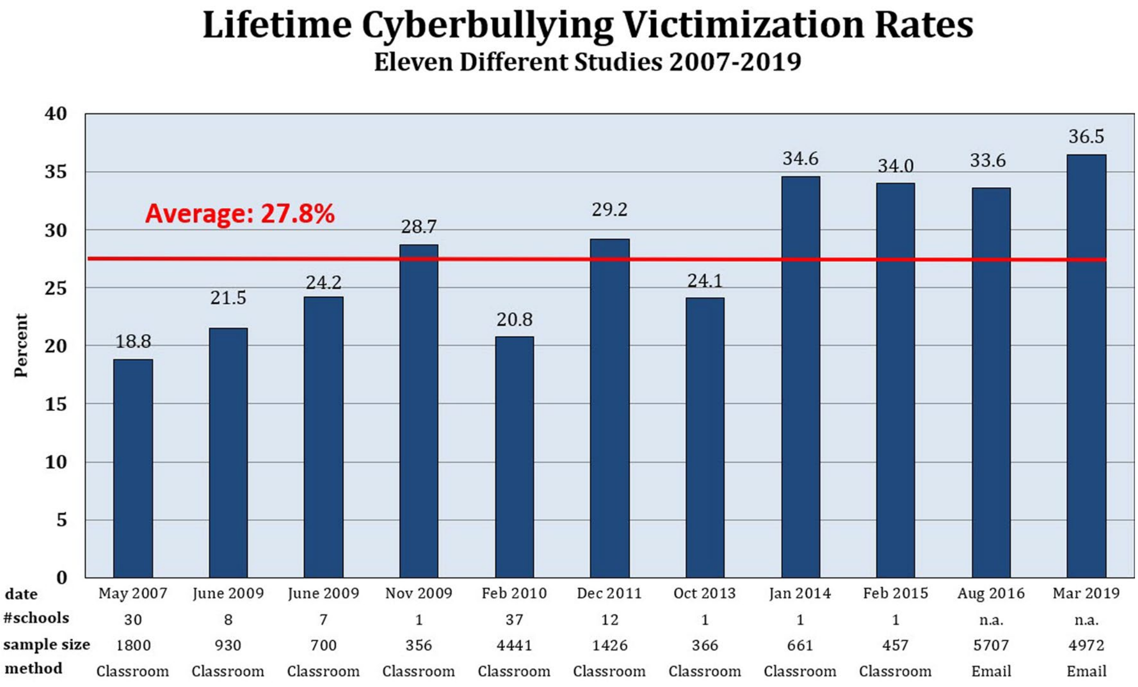
[1] Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

The rest of this paper is systemized as follows. Section II briefly discusses the motivations for this study. Section III shows some background information about Cyberbullying and Cyberbullying detection approach. Section IV describes the dataset and supporting tools that are used to construct the bully detection model on social media. Section V summarizes and analyzes the existing literature closely related to our study. Section VI discusses and addresses the main issues and research challenges associated with building the cyberbullying classification systems, and the survey conclusion is discussed in Section VII.

## 2 Motivation

According to a new poll [4] released by UNICEF on September 3, 2019, one out of three adolescents in 30 nations said they had been victims of online cyberbullying. One out of five reported that they had been skipping school due to online harassment and violence. Nearly half of the young generation (47%) [5] has received harassing, threatening, and nasty messages through OSM. Justin W. Patchin and Sameer Hinduja [6] have surveyed their cyberbullying research center. They have collected data from middle school and high school across the USA in eleven unique projects from 2007 to 2019. They reported that, on average, 28% of students who are part of their project claimed that they had been victims of cyberbullying at some point in their lives. The average rate of Cyberbullying victimization is shown in Fig. 1.

Symantec [7] concluded that around 80% of people in India are victims of cyberbullying. Almost 63% are victims of inappropriate words and shaming. 59% are the victims of unseemly rumors about them degrading their social persona. The exponential growth of online communication technology and social media websites have worsened. Many studies have also shown that people who have faced cyberbullying are prone to a higher risk of suicidal tendencies [8, 9]. Some other studies also reported the relation between cyberbullying victims and the risk of suicidal ideation [8, 9]. For example, Natasha MacBryde, a 15-year-old schoolgirl, committed suicide after getting harrying comments from anonymous users on her Form spring social network and being called "Slut” by her school friends [10]. Hannah Smith, a 14-year-old girl, also committed suicide after being cyberbullied on the Ask.fm social network [10]. With all these considerations, this study has been carried out to compare the predictive abilities of the existing models for cyberbullying on social media websites using ML and DL and highlight the factors that challenge their performance and efficacy.



Justin W. Patchin and Sameer Hinduja
Cyberbullying Research Center
www.cyberbullying.org

**Fig. 1** Rates of cyberbullying Victimization [6]

# 3 Background

## 3.1 Cyberbullying

Cyberbullying is a fraudulent activity online via electronic gadgets and can occur on various platforms where people share content, activities, information, and interest. In another way, cyberbullying is defined as an activity in which using the internet, mobile phone, video games, or any digital technology and one may post or send text, images, or videos to embarrass or humiliate someone or a group of people. Examples of cyberbullying messages are as follows: threatening messages, revealing personal information to other people without consent, sharing and forwarding the private message to others, spreading rumors via text messages or online social accounts to make people gossip, etc. Our review [11–16] suggests that cyberbullying attacks may be classified into ten categories. These are masquerading, flaming, trolling, flooding, harassment, cyberstalking, denigration, exclusion, outing, and trickery, as shown in Fig. 2. and explained as follows:

- *Masquerading*: A malicious user pretends to be someone other and posts harmful information to get the victim in trouble or menace or harm that person's status [12, 14].
- *Flaming*: It is a kind of online fight that involves insulting messages, abusive words, or flames between the users [12, 14, 15].
- *Trolling*: When someone creates conflicts on social media sites to create a controversial or inflammatory environment to provoke an emotional response from the users [16].
- *Flooding*: It is another type of bullying where the felon frequently sends the same message or comment to the victim who does not participate in the conversation [11].
- *Harassment*: In this scenario, the victim repeatedly receives insulting, rude, and offensive messages [12].
- *Cyberstalking*: It is another form of offense where delinquents gather personal information on the victims, spreading false rumors, threatening harm through email, etc. [12].
- *Denigration*: It is also known as "defamation"; it is an abusive attack on a person's character or his reputation using untrue statements [12, 14].
- *Outing*: Posting someone's private or awkward information, images, or video in a public chat room or environment [12].
- *Exclusion*: Such type of bullying generally happens between teenagers and youth. They deliberately exclude someone from an online group [13].
- *Trickery*: It is a deceptive activity where people use the trick to cheat others [14].

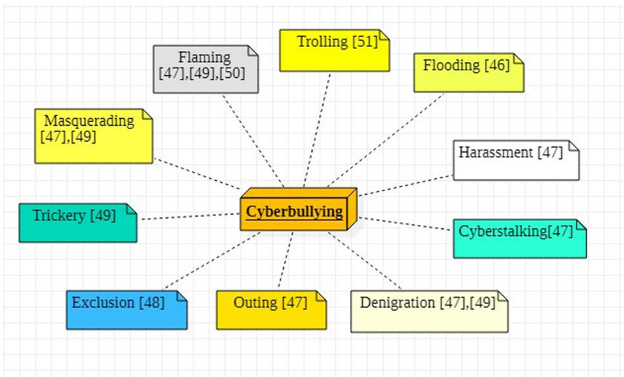


**Fig. 2** Common Forms of Cyberbullying

Cyberbullying is even more devastating for weak-hearted people prone to depression and has adverse physical and mental health effects. Thus, cyberbullying is a pressing social problem that needs to be controlled urgently and constructively.

Where does Cyberbullying transpire?

Cybercrime or cyberbullying has increased with easy access and extensive use of technology. Crime mainly arouses child development, public security, and an adult's socio-economic conditions [17]. It is challenging to figure out specific platforms that may or may not be prone to cyberbullying attacks. There is no way to ensure which one is more secure or unsafe than the other. Cyberbullying may materialize virtually on social media (through Instagram, Twitter, Facebook, WhatsApp, Snapchat, etc.), Email, YouTube, Online multiplayer games, etc. Statistics show that Instagram is the most popular platform for cyberbullying, where the victimization rate is relatively high and closely followed by Facebook and Snapchat [18].

The majority of youth aged between 10 and 17 years have been reported harassed or bullied over the internet and on social media. 42% of adolescents experienced cyberbullying on Instagram, the second was 37% on Facebook, and the third was 31% on Snapchat, while YouTube and Twitter were running closely behind at 9% (Fig. 3).

## 3.2 Cyberbullying detection approach

It is arduous to detect, prevent and control cyberbullying because of the misuse of the internet and technology. Most researchers use standard techniques to identify cyberbullying instances based on Social Network Mining. It is a combination of Social Network Analysis and Data Mining.

A general approach to designing a predictive cyberbullying model consists of data collection from relevant sources like OSM, preprocessing these data, and extracting and

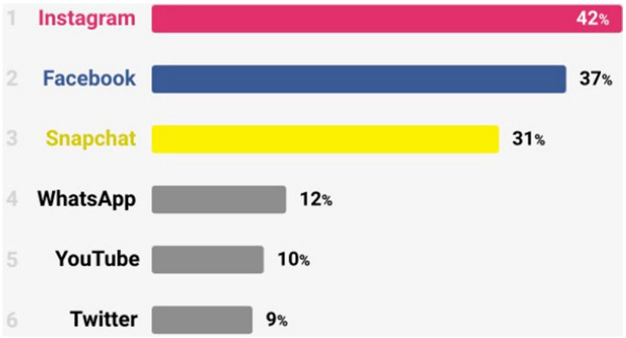


**Fig. 3** Cyberbullying on Social Media Platforms [18]

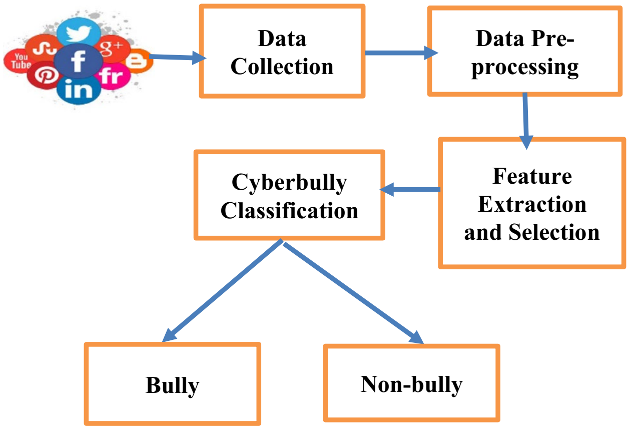


**Fig. 4** Cyberbullying Detection System

selecting valuable features. Ultimately, develop a cyberbullying classification algorithm to identify whether a person is a bully or not in cyberspace. A diagrammatical representation of the cyberbullying detection system is shown in Fig. 4.

### 3.2.1 Data collection

Designing a reliable cyberbullying detection model depends on the collected dataset. Data collection from social media is a challenging task because of its privacy. Many researchers draw out the data from social media by specific keywords or profile IDs, but searching posts with these particular keywords may yield a biased result [19]. On the other hand, the authors extracted the data from social media using public social media APIs. Extensive research faced quality issues in data provided by an API, such as a lack of clarity regarding biased results, insufficient documentation of data, etc. Social media public APIs are another impediment, providing an inadequate amount and data type. Data can be in the form of text or visuals. Generally, scholars use the social media public API or web scraping tool for Data extraction in this area. They extract the data with specific hashtags like #bully, #cyberbully, #bullying, #cyberbullying, #stopbullying, #antibullying, #bullyinghurts, etc. Most of the study in this area has been processed on publically available tweets (Twitter dataset).

### 3.2.2 Data preprocessing

Data Preprocessing is the next crucial step while building models, especially while working with text data. The performance of ML models depends directly on how well the data have been preprocessed before feeding it to the model because the text is transformed into a more consumable form during data preprocessing. Thus, the ML algorithms perform even better. Data are preprocessed in the following ways:

i. Data cleaning—It removes the special symbol, numbers, URLs, etc.
ii. Tokenization—It separates each post into multiple small pieces called tokens, i.e., larger chunks of text samples are divided into paragraphs, paragraphs are split into sentences, sentences are divided into words, and so on. Tokenization is used because it is impossible to feed the whole text sample simultaneously to the model. Also, the meaning of the text remains the same and can easily be interpreted using the words present in the text.
iii. Transform Case—It transforms lower case to upper case and vice versa to differentiate between upper and lower case letters.
iv. Stemming/Lemmatization—It refers to reducing a word to its base word. Stemming uses the only stem of a word, while lemmatization analyses a broader aspect of what a word means in a sentence by just recognizing its root word. Stemming may be faster than lemmatization. But in the case of accuracy, lemmatization is more effective.
v. Encoding Labels—Encoding labels refer to converting output labels into a numeric form so ML algorithms can operate and use them better. For example, an attribute having output classes angry, sad, hurt, etc. In the encoding label, angry is replaced with a number, say 0, sad is replaced with 1, hurt is replaced with 2, etc.
vi. Removing Outliers—Outliers are observations in a dataset that do not conform, i.e., they have an excessive deviation from the other observations of the dataset. The prevalent type of outliers is the observations far from the rest of the observations or the center of mass of observations.
vii. Padding—All the ML models require inputs to have a similar shape and size. Thus, when preprocessing the text data, we need to pad them before using them as inputs in the model, as the sentences may not naturally

have the same length. Accordingly, researchers padded the short sentences to that length and truncated the longer ones. Various automated tools are available for data preprocessing like WEKA, WordNet, etc.

### 3.2.3 Feature extraction

After data preprocessing, the system extracts the most befitting feature from the preprocessed text. Preprocess posts are examined using standard feature extraction methods like Bag of Word(BOW) [20, 21], TF-IDF (Term Frequency-Inverse Document Frequency) [22–24], Word2Vec [25], N-gram [22, 24, 25], bigram [26], etc. and transformed into a vector space model where posts are represented as a count vectorizer form. Some of the standard feature extraction techniques are discussed in brief as follows:

- *Bag of Word(BOW)*: The bag of words (BOW) model extracts the feature matrix. In this model, the text is represented as a corpus or multiset of the contained words, not considering word order and word structure but retaining multiplicity. It is a representation describing the occurrence of words in a text. Each word is represented as a feature, so all unique words constitute a feature set. The idea behind this model is that two pieces of text are similar if they have identical content. The BOW model is inefficacious for a large number of datasets. If we have a large corpus, it has many unique words. So table size is enormous, and to process it, the model takes a significant amount of time, i.e., high time complexity.
- *TF-IDF (Term Frequency-Inverse Document Frequency)*: TF-IDF is a statistical metric that indicates a word's importance in a document. It is a knowledge-gathering technique that measures a word's frequency (TF) score and inverse document frequency (IDF) score of each word that occurs in the document. The multiplication of the TF and IDF scores of a word is called the TF-IDF score. Choose simply the higher the TF-IDF score (weight), the rarest the word is in a given document, and vice versa. The TF-IDF score converts each word into a fixed-sized vector and feeds them to an ML classifier. The TF-IDF feature extraction techniques improved the performance of classifiers in [22, 24, 25].
- *Word2Vec*: Word2vec is a NLP technique. This algorithm uses a neural network model to find the semantic relationship between the words. The function of word2vec is to convert each word into a fixed-size vector space and capture the similarities between the words. This model utilizes the two-way architecture; Continuous Bag of Words (CBOW) or skip-gram. CBOW predicts the target word using the context word, while the skip-gram model predicts the context word using the target word. Initially, it builds the vocabulary from the input and converts it into vector form. Then it calculates the cosine similarity value and determines how similar they are. Words with similar contexts are located closer in the vector space [27].

  *N-gram*: N-gram is one of the feature extraction methods which turns the word into vector form. It is the extended version of BOW because it is more informative and captures the semantic relation around each term [22]. An n-gram has a sequence of n words. A sequence of two words is called bigram(2-gram), and of three words is known as trigram(3-gram), "welcoming all" or "all emails" are the example of bi-gram, and "this is awesome" or "Oh My God" are the example of tri-gram. Combining characters or words bring out meaningful information; therefore, N-gram plays a crucial role in cyberbullying detection. Occasionally it may also be meaningless and tend to create noise. Zhang et al. [25] employed the principal component analysis method that applies an orthogonal transformation to convert possibly correlated features into linearly uncorrelated features to resolve this type of noise.

### 3.2.4 Feature selection

After performing the feature extraction, the feature selection methods play a crucial role in the legitimate performance of the classifier. The primary aim of the feature selection algorithm is to select the vital features without minimizing the prediction rate of models. Chi-Square (CHI2), Information Gain (IG), and Pearson Correlation are the typical feature selection method used in offensive content detection. These methods significantly filter out less feasible features to improve classification accuracy. These standard feature selection methods are briefly explained as follows:

- *Chi-Square (CHI2)*: This well-known statistical method evaluates the relation between sample classes. It measures how the expected frequency and observed frequency of two features differ. If the difference between the expected and observed values is high, it signifies a high chi-square value. The features with higher chi-square values depend on response, which means more relations between the class and test features. Thus, these top features are selected for further processing. According to [28], authors performed the CHI2 test and found that retweets, favorites, sender location, and sender followers highly depend on social media features.
- *Information Gain (IG)*: It utilized the Decision Tree algorithm to select the feature and measure each node's entropy for decision-making. IG signifies the complete information provided by a feature item for a class. The higher the IG, the greater the impact of that feature in determining the classification of the document. Gener-

ally, features with the highest information gain score are considered efficient classification.

- *Pearson Correlation*: Usually Pearson Correlation feature selection method is used in dimensionality feature reduction to estimate optimal feature. This method finds the linear correlation between the actual class and extracted class. The correlation value lies between − 1 and + 1. − 1 value represents the absolute negative class (if one feature increases, the other decreases, and vice versa). + 1 value represents the absolute positive class (if one feature increases, then the other feature also increases and vice versa), and 0 represents the no correlation between the two features. If the correlation between the two features is high, i.e., they are more linearly dependent, we may drop the one feature of the two.

### 3.2.5 Cyberbully classification

ML is a set of methods used to detect and find the pattern and relationship in the data [29]. ML algorithms consistently improve themselves with experiences and past data. The system learns the pattern and predicts the class (Cyberbullying and Non-cyberbullying). There would be a corpus of labeled data and unlabeled data for cyberbullying detection, based on which we may need to use supervised or unsupervised learning paradigms. However, labeled and unlabeled data could be used with the semi-supervised methodology used for cyberbullying detection in OSM [30]. In this way, ML handles the noisy and un-uniformed data, where some data components are irrelevant or less important for the decision functions.

Researchers primarily emphasize deep learning in this era, but we concluded that ML and DL are both essential in our studies. The selection of models depends on the size of the data, data type, the complexity of the problem statements, and the system's configuration. Based on our studies, we have categorized the classification model into Machine Learning-Based Model and Deep Learning-Based Model.

a. Machine Learning-Based Model

Following are the standard ML methods which have been used frequently for cyberbullying detection explained in brief:

- *Naïve Bayes (NB)*: It is a supervised learning algorithm based on conditional probability with an assumption of no dependency among features. Based on the training data metrics, the model calculates the likelihood of belongingness to a particular category with the highest posterior probability. NB is a frequently used ML algorithm dedicated to tasks like Text classification, Real-time prediction, multiclass prediction problems, etc. Nalini and Sheela used the Linear Discriminant Algorithm with NB probabilistic method to construct a sentiment classification for Twitter's cyberbullying messages [31]. NB classifier was also applied with text mining to detect twitter users' emotions with 83% accuracy [32]. Many examined studies used the NB algorithm to construct the cyberbullying prediction model in [26, 33–35], and [36].
- *Support vector machine (SVM)*: SVM is also a supervised ML paradigm generally used for Text classification [37]. It is a statistical classification algorithm aiming to correctly identify a hyperplane in N-dimensional space that segregates the data points. It generates a hyperplane based on the feature attributes of each category. The distance between the hyperplane and the nearest features of each type is maximized. It uses support vectors that are nothing but the vectors for every class closest to the hyperplane to calculate the hinge loss. It emphasizes reducing the misclassifications to achieve minimum classification risks. SVM also supports many kernels like Radial basis function(RBF), Linear kernel, Polynomial kernel, Sigmoid kernel, and Gaussian Kernel. It also supports the classification of nonlinear relations as well. The advantages of SVM are its scalability, reliability, high speed, the capability to classify in real-time environments, and dynamically update training patterns. Chen et al. [33] used SVM and NB to build the cyberbullying model for detecting destructive and offensive language on Social Media. The result showed that SVM is more efficient and accurate than NB, but NB is faster than SVM. The SVM techniques enhanced the prediction of cyberbullying in OSM. Andriansyah et al. [38] constructed an SVM-based prediction model to see how far it could correctly classify a comment on Indonesian accounts as cyberbullying or not.
- *Logistic Regression (LR)*: Regression analysis needs to quantify data, and it constructs the separated hyperplane between the two classes by employing the logistic function. LR is relatively similar to the neuron. It is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression, which outputs continuous number values, LR converts its output with the help of the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. The sigmoid function is calculated based on Eq. (1).

$$P = \frac{e^{B_0+B_1x}}{1 + e^{B_0+B_1x}} \tag{1}$$

where $B_0$ and $B_1$ are bias and coefficient values, respectively, and $x$ is the independent variable.

When features are indefinite, we must change them to numeric values for regression analysis. For example, we can represent a social media post for bullying as 1 and non-bullying as 0. Suppose the probability is close to 1; in that case, observation is likely to be part of the bullying class. If the probability is close to 0, then observation is a part of the other class. These studies used LR to construct the cyberbullying classification model [22, 26, 39, 40].

*Decision Trees (DT)*: DT is a nonparametric supervised learning method used for classification and regression. It is a simple tree-like structure, and the model decides every node. It is helpful in simple tasks. It is one of the most popular algorithms. It has various advantages like easily explaining ability easy showing how a decision process works. It classifies the raw data in the dataset by dividing and recursively solving the root's problem statement until it reaches the leaf nodes (terminal nodes), representing a particular class. The most common version of a decision tree used to construct the cyberbullying model is C4.5, the advanced version of ID3. The tree's size and the model's accuracy are the essential factors of the prediction model [41]. A complex and large tree may be over-fitted, and a small tree may specify that the training set is unbalanced. DT has been used to construct cyberbullying models to predict aggressive and offensive language threats in OSM [34, 35].

*Random Forest (RF)*: RF is a popular and straightforward ML algorithm that belongs to the supervised learning data mining technique. It can be helpful in both Regression and Classification problems in ML. It uses ensemble learning; it combines different classifiers to resolve a complex issue and improve the model's performance. It is an advanced form of the standard decision tree model and contains several decision trees on various dataset subsets. It takes the mean of the data to improve the predictive accuracy of that data set. Instead of depending on one decision tree, the RF takes the prediction from each tree and predicts the final output based on the majority votes of forecasts. This methodology has also been used to construct reliable cyberbullying prediction models in [23, 42, 43], and [44].

*K-nearest neighbor (KNN)*: KNN is a nonparametric and instance-based classification paradigm. It mainly uses Euclidean distance as the distance metric [45]. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance. In KNN learning, the target function may be for real or discrete value. The main task of KNN is to classify the unknown instances by computing the distance or similarity distance between the training and testing dataset. It is mainly used for categorization extrapolative troubles in engineering. The following two properties could explain KNN in good health.

- Lazy learning algorithm – KNN is an indolent knowledge algorithm since it does not contain a particular preparation stage plus uses every one of the facts for preparation as categorization.
- Nonparametric learning algorithm – KNN is also known as a nonparametric knowledge algorithm since it does not take for granted the fundamental data.

KNN algorithm is an efficient and straightforward classification technique that has been used for text categorization [45]. These studies efficiently used the KNN algorithm to build cyberbullying classification models on social media [34, 35].

Many ML algorithms exist, but the techniques mentioned above have been more frequently used to efficiently predict cyberbullying on social media. So far, most of the research on cyberbullying has focused on supervised learning. There are multiple challenges [46]. Our review found that SVM is the established classifier used for cyberbullying prediction and is closely followed by NB.

b. Deep Learning(DL)-Based Model

Deep learning-based methods have gained immense popularity in image, speech, and text analysis fields. DL is a subset of ML that employs a deep neural network. DL mainly comprises interconnected layers. These layers imitate work as human neurons, where each layer is connected to resemble the human brain's functionality. The deep network generally consists of at least three layers, an input layer, an output layer, and one hidden layer. In DL, each layer learns and transfers its output as input to the next layer to compute the correct result. Various deep neural networks like Convolutional Neural Network(CNN), Recurrent Neural Network(RNN), Long Short-Term Memory(LSTM), Bidirectional LSTM(BLSTM), etc., are used to detect cyberbullying [47, 48]. The deep neural network applies to images, as well as text. In an image, only vector pixels are replaced by a vector of words, and the rest of the process and structure remain the same. A deep neural network was used to construct the cyberbullying prediction model in [48–52]. It can predict the result in a more intelligent way rather than a traditional classifier.

Following Deep Learning-Based models are used in the construction of cyberbullying detection systems.

- *Convolutional Neural Network (CNN):* CNN is immensely used in Machine Learning, Computer Vision, and Pattern Recognition and proved extremely useful with excellent results. CNN learns from the hierarchy of patterns in data and gathers the most complex patterns with the help of hidden layers. It breaks the complex pattern into smaller patterns and then learns it accordingly. Convolutional neural networks are based on convolution property to make the prediction. They are predominantly used for image processing because of their 2D nature. They generally consist of 2D filters, easily detecting different image edges and styles. Using 2D filters also reduces the number of parameters per unit and thus the computational needs. For NLP tasks, CNNs are often utilized in 1-dimensional forms. In one-dimensional CNN, the filter or the kernel sides along one data dimension (like a moving window). In the case of text-based data, the fixed-sized word vectors are used; thus, the sliding dimension is the sentence or the flow of the text. These studies [24, 48–50, 53] used CNN to construct cyberbullying detection model efficiently.

- *Recurrent Neural Network (RNN)*: RNNs, a class of Artificial Neural Networks, are employed to identify data sequence patterns such as spoken words, text, numerical time-series data, etc. In RNN, information flows across hidden layers. Hidden layers provide the capacity to recall information previously processed. This feature gives an edge for processing the time-series or sequential data. Because they have memory, dynamic networks separate themselves from static networks. The existence of memory components in complex networks makes them a possible candidate for sequential data processing. Therefore, the output of a hidden node at the moment depends not only on the current input of the node but also on the output of the same node at the previous time. The valuable information is stored in hidden nodes and can be utilized later. RNNs are repetitive as the same task is performed for each sequence characteristic, and the output relies on the preceding output. In RNN, the hidden state behaves as memory and maintains the relations in a sequence. The cells in RNN share the same loads throughout the time. The output $O_t$ and hidden state $h_t$ of RNN at time t are given as:

$$h_t = \sigma\left(W_h h_{t-1} + W_x x_t + b_t\right) \tag{2}$$

$$O_t = \sigma(W_s h_t) \tag{3}$$

where $\sigma$ denotes a Sigmoid activation function, $W_h$ represents the weight matrix between the previous hidden state and hidden layer, $W_x$ represents the weight matrix between the current input and hidden layer, $b_t$ denotes bias at time $t$, $h_t$ represents the hidden state at time $t$, $x_t$ denotes the input of network at timestamp $t$, and $W_s$ represents the weight matrix between the hidden layer to the output layer.

RNN is facing the "Vanishing Gradient problem"[54]. As a result of this problem, it is challenging to learn the initial layers, and the predictions made are incorrect. Because the initial layers cannot think, it is difficult to recall the data of initial time stamps and suffer from a short-term memory crisis. In simpler terms, if the gap between the related knowledge and the position where the meaning of that information is required is significant, then because of the short-term memory problem, RNN fails to provide detailed results. Cyberbullying detection model constructed using RNN in [50, 55].

- Long Short Term Memory (LSTM): This model is based on RNN network architecture that is extensively used in areas of deep learning. Earlier used neural networks did not have the feedback mechanism, whereas these models have feedback connections. It can run a single data point at a time but can process many sequences simultaneously. The specialty of this model is that they have memory power which remembers the past inputs. They are the building blocks for the whole layers of RNN architecture and overcome the RNN's problems (gradient-vanishing/short-term memory) [54]. LSTM overcomes these problems by using a gradient-based learning algorithm which enforces a consistent error flow across the internal states of the LSTM cell. As the error is constant, the error neither vanishes nor explodes. LSTM uses multiplicative units(MUs) to prevent the error from unwanted vanishing; that is, constant error flows can be obtained by using MUs [48]. In the memory blocks of LSTM, there are three components.

  *Forget Gate*: The gate decides which input must be replaced with the new information in this cell section. To retain inputs, it gives a value close to 1, forgetting the input. It provides a value close to 0.

  *Input Gate*: Based on past learning and inputs, it decides which information should be stored.

  *Output Gate*: The various inputs and the cell state decide which information will be passed to the next phase.

  LSTM is efficiently used in these studies [48, 50, 55].

  *Bidirectional LSTM (BLSTM)*: This model is also based on RNN network architecture. The extended version of LSTM will run the inputs in two ways (Backward and Forward), one from past to future and one from future to past. Thus, this LSTM, using the two hidden states combined, can preserve information from both past and future at any time. The idea behind the bi-directional network is to collect more knowledge about the input data. A BLSTM will feed the following letter in the sequence on the backward pass to access future information. Unlike the traditional LSTM structure, BLSTM consists the two different LSTMs that run on sequential data. The BLSTM model improves the performance of the sequence classification process. Initially, BLSTM was only applied in some specific areas, but now it has gained more popularity and is applied to various fields like speech recognition, face detection, text classification, etc. BLSTM with neural attention (BLSTM with attention) model [48] captured the text's semantic information without using the NLP system or any effort and showed excellent performance in cyberbullying detection. These studies [55–58] used BLSTM to build cyberbullying detection system.

#### 3.2.6 Performance analysis

Many factors may contribute to the result. Most authors have considered the Accuracy, Precision, Recall, and F-score measurement as a reliable measure for evaluating cyberbullying predictive models' efficiency [22, 25, 26, 59–61].

- *Accuracy*: This performance metric measures the correct classification (Bullying and Non-bullying). It is formulated as:

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{4}$$

- *Precision*: This performance metric measures the correct classification of cyberbullying posts. It is formulated as:

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{5}$$

- *Recall*: This metric performance measures how the algorithm identifies many bullying posts. It is formulated as:

$$\text{Recall} = \frac{T_P}{T_P + F_N} \tag{6}$$

- *F-Measure*: It is a harmonic mean of precision and Recall. The following formula computes it:

$$F - \text{Measure} = \frac{2(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \tag{7}$$

where $T_P$ = True positive value (Total number of correctly, identified offensive words by the classifier). $F_P$ = False-positive value (Total number of incorrectly identified offensive words by the classifier). $T_N$ = True negative value (Total number of correctly identified non-offensive words by the classifier). $F_N$ = False-negative value (Total number of incorrectly identified non-offensive words by the classifier).

## 4 Dataset and tools

The first step of any research is data collection, and a dataset is useless until information, knowledge, and implications are extracted. A large amount of actual data is required for training and testing purposes. Data collection from social media is a very challenging task. It requires specialized tools and techniques to collect, analyze, and build a valuable repository for it. Many open-source resources are available to obtain massive datasets for analysis and learning, like UCI Machine Learning Repository, Kaggle, sentiment140.com, SNAP, etc.

One may build the dataset using social media public API or extract the dataset by using some specific keywords or profile IDs. In another way, data can also be collected from the World Wide Web by manual copying or web scraping tool. In our study [20, 42, 62–71], we recognized some standard tools and techniques are followed for data acquisition and preprocessing. These are discussed as follows:

Content Analysis on the Web 2.0(CAW 2.0) is a web mining technique that understands and discovers the information on the web. CAW 2.0 provides the datasets for misbehavior detection [71]. The authors efficiently detected the harassment activities on Web 2.0 using the document's content, sentiment, and contextual features by CAW 2.0.

NLP requires a unique supporting tool such as WordNet [66], an extensive lexical database first proposed for the English Language at Princeton University. It contains only nouns, verbs, adverbs, and adjectives. It provides the similarity between the words used for NLP applications, for example, word sense disambiguation, representations of contents of a document, and term expansion in information retrieval systems.

Waikato Environment for Knowledge Analysis (WEKA) tool is also used for data analysis. It is an open-source data mining tool that consists of the ML algorithm and contains the tools for data preprocessing, classification, regression, clustering, association rule, and visualization [68].

Twitter Application Programming Interfaces(API) is a tool used to extract the data from tweets. There are three twitter data interfaces available for extracting the data. Streaming API (stream the real-time tweets and required continuous internet connection), REST API (Retrieve the only query-related information), and Search API (search only simple queries) [65].

Amazon Mechanical Turk (AMT) is a web service used to label data. AMT is an online crowdsourcing marketplace that enables individuals or businesses (Requestors) to post tasks (Called human intelligence tasks), which paid workers often do. The requestors pay workers as per their completed Human intelligence task(HIT). This tool's primary goal is to label the post as either cyberbullying or not [69, 70].

Grasso et al. [62] introduced a new scraping technique rather than complex scripting, heavyweight, or visual tools. They use OXPath, which is a minimalistic wrapping language. It turns into two simple tasks: first, identify the relevant node through the expression of OXPath, and second, specify what action to perform on those nodes.

Van et al. [42, 63] collected the Dutch language posts (91,370) from the OSM site Ask.fm, where anonymous users can create a profile to ask questions and answer them. They extracted the data using the GNU Wget software. After some refinement of non-Dutch data, it resulted in 85,462 posts.

Haidar et al. [64] collected Arabic Tweets from Twitter in the Middle East and Gulf region. They scrapped the Arab tweets around the 100 km region of Arab. After the scrapping, the Twitter content, refinement, or preprocessing techniques were applied to remove non-Arabic literals, Hashtags, and retweets. After preprocessing, word embedding techniques were used for data preparation.

According to [20], the authors developed a java program to extract social media data (Facebook, Twitter). Facebook Graph API is used to collect data from Facebook, and Twitter REST API is used for Twitter. 2400 Bangla text contents were collected, and bag of word (BOW) techniques were used to train the prediction model.

State-of-the-art techniques and supporting tools for data preparation are shown in Table 1.

## 5 Related work

This section reviews the existing techniques to construct a cyberbullying detection system and discusses some insightful information the authors have addressed. It is difficult for scientists and researchers to identify cyberbullying because of its ambiguous nature and its categorization. Patchin and Hinduja defined "cyberbullying as a willful and mischievous activity, which repeated harm inflicted through the medium of capturing electronic text" [13]. Such mischievous activity can be controlled by either software filtration methods or ML techniques. Social networking sites employ a software filtration method that automatically recognizes the bully's post or comment and automatically deletes or shades the bully's words [14]. Tibor et al. [72] approached another alternative method to detect cyberbullying by comprising normative agents in a virtual society that controls and regularly monitors the active user within a virtual community. Based on BDI-Model, the normative agents use several methods and techniques to identify the various forms of violation and aggression that happen exclusively via the online medium. The working principle of all processes is based on ML or DL techniques. The main issue with these techniques is that some classification techniques efficiently work only on textual data, while others efficiently work on only multimodal data. Thus, we categorized our study into two subsections: Identifying Cyberbullying on Textual Data and Identifying Cyberbullying on Multimodal Data. We also discussed every aspect of these two subsections, which will be helpful for the newcomer to work in this field.

### 5.1 Identifying cyberbullying on textual data

Nowadays, social media is the most popular platform where people are extensively affected by the problem of bullying. The form of cyberbullying content on social media is either textual or multimodal. Examining cyberbullying content on textual data is always challenging for humans or machines. Initially, researchers used simple classifiers and hard-coded features. But later, ML gained popularity in text classification problems, which provides the right direction for researchers to minimize such bullying problems. Reynolds et al. [41] performed KNN, SVM, DT, and Rule-based Jrip classifiers on a dataset collected from Formspring.me, a Q&A formatted website containing various bullying textual data. The authors labeled the dataset using AMT and predicted the cyberbullying content with 78.5% accuracy, but their model performs well on only a small amount of data. Chen et al. [33] proposed the Lexical Syntactic Feature (LSF) new architecture using the SVM, based on a language model to automatically detect offensive content and offensive users on OSM.

LSF architecture deals with the messages and examines the user's witting style, structures, posting patterns, and post content. They applied their model to large datasets (YouTube comments), tested different evaluation metrics, and showed

**Table 1** State-of-the-art data preparation tools and techniques

| Refs | Dataset source | Dataset extraction technique | Supporting tool |
|---|---|---|---|
| [41] | Formspring | CAW 2.0 | Amazon's Mechanical Turk |
| [47] | Facebook, Twitter | Facebook Scrapper, Twitter Scrapper | WEKA |
| [69] | ASK.fm | Scrapy Crawler | WordNet |
| [100, 101] | Instagram | Instagram API | Amazon's Mechanical Turk |
| [102] | MySpace | Barcelona Media | WEKA |
| [103] | Twitter | CAW 2.0 | WEKA 3.0 |
| [36] | Twitter | Twitter streaming API and Twitter4j | Amazon's Mechanical Turk |

excellent results with running time advantage. Mangaonkar et al. [26] focused on the collaborative paradigm over the stand-alone paradigm to improve detection. In a collaborative cyberbullying detection system, each entity acts as an autonomous detection node, and the collaboration of these nodes helps the system classify the bully content. The authors used three different collaboratives; heterogeneous collaboration, homogeneous collaboration, and selective collaboration. They recommended that the heterogeneous collaborative approach performs well with running time advantages. Chavan et al. [22] improved the classifier efficiency using two additional features; capturing pronouns and using the skip grams method to create a feature vector. These features increased the accuracy by 4%. Al-garadi et al. [73] handled the imbalanced dataset using the synthetic minority oversampling technique (SMOTE). The proposed new features such as network, activity, user, and Twitter textual content improve the classifier's performance. Twitter users' psychological features like personalities, emotions, and sentiments also increase the model efficiency [32]. Sometimes multiple textual features are used to construct an optimum cyberbullying detection model [25].

With the extensive use of technology, emoticons have gained popularity these days. Sometimes, OSM users express their opinion on OSM by emoji or text with emoji. Emoticons directly express sentiment, and negative sentiment expression always leads to cyberbullying. The polarity of sentiments represents the sentiment expression; it may be positive or negative. Thus, the post with emoticons directly depends on sentiment polarity [60].

Cyberbullying is not limited to only English; it may also happen in other languages like Hindi, Bangla, Turkish, Arabic, etc. Sometimes, offenders also use multilingual or bilingual languages; thus, it has always been challenging for the researchers. Much research has been proposed for cyberbullying detection in the English language. However, some work has been done to address cyberbullying detection in multilingual environments or other languages. Eshan et al. [23] detected the Bengali abusive text on OSM. They performed various classifiers like RF, Multinomial NB, SVM with linear radial basis functions, polynomial, and sigmoid kernel functions. The result showed that SVM with Linear Kernel algorithm with unigram, bigram, and trigram-based TfidfVectorizer performs better than CountVectorizer, and Multinomial NB performs well in most of the cases. Haider et al. [47] detected the cyberbullying content over Twitter using SVM and NB classifier and applied the TweetToSentiStrengthFeatureVector to convert string to a word vector. Their work proved that detecting bully content in the Arabic language is also possible, but the performance of their model was not satisfactory. Pawar et al. [59] used the synthetic data generation technique to solve the class imbalance problem in the multilingual cyberbullying detection process. Synthetic data generation generates additional instances of bully content from the existing instances; it could help them increase the classifier's performance.

The Deep learning-based model is also an effective technique for detecting cyberbullying and has improved Recall and precision than the baseline ML algorithms [74]. CNN can be applied to text and multimodal data, especially on images. Al-Ajlan et al. [49] employed the optimized Twitter cyberbullying detection based on deep learning(OCDD), a DL model for cyberbullying detection. Unlike other conventional techniques, it eliminates introducing new features in the dataset. As no features get added to the data, steps like feature extraction and selection get untangled; otherwise cumbersome to perform. Rather than extracting features from a tweet, OCDD vectorizes it, thus preserving the exposition of the word. The Glo-Ve technique was used to produce word embedding that outperformed other techniques. The word vectors were then fed into CNN for classification, and a metaheuristic optimization algorithm was used to decide the appropriate number of parameters. OCDD classifies Twitter posts more efficiently and intelligently than traditional methods. Chu et al. [50] presented their idea to flag offensive comments using the DL models. They tested three models: RNN with LSTM, CNN with word embedding, and CNN with character embedding. They compared their results and argued that using DL models gives better performance than non-DL models. Most of the authors only focused on one of the following bottlenecks; either they considered only one particular OSM, or they focused only on one topic of a cyberbully, or they relied only on the manual labeled feature. The authors showed that the deep learning-based model overcame all three bottlenecks [48]. The authors have collected the data from Formspring. me, Twitter, and Wikipedia. They experimented with deep neural network-based models (CNN, LSTM, BLSTM, and BLSTM with attention). BLSTM, with attention, achieved the highest *F*-score of 0.94 for two datasets out of 3 datasets. Iwendi et al. [55] compared 4 DL models—Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN)—to determine their effectiveness in cyberbullying detection. The proposed LSTM used doubled input gates, output gates, and forget gates, known as BLSTM, which obtained the highest accuracy (82.18) and outperformed other models. Kumar et al. [56] proposed a multi-input integrative learning model based on deep neural networks(MIIL-DNN) to tackle cyberbullying detection in real time. The research concentrated on code-switching data, specifically Hinglish. The model took English, Hindi, and typographic features as input learned through capsule network, BLSTM, and multi-layer perceptron(MLP). The model proved advantageous as it did not increase the dimensionality of the input data. The model classified text into a broader spectrum of bullying and non-bullying compared to previous research that was rather parochial. A splendid ROC-AUC of 0.97 was obtained on datasets scrapped from Twitter and Facebook.

Character detection as a social media bully is still a challenge for short, noisy, and unstructured with spelling errors. Lu et al. [24] proposed Char-CNNS (Character-level Convolutional Neural Network with Shortcuts) to detect cyberbullying on OSM. The Characters were used as the basic unit of learning to tackle the hurdle of spelling errors and deliberately attempt to create confusion. Shortcuts were used to integrate various levels of features so that the model could learn every minute bullying signal. A focal loss function was adopted to ensure that the class imbalance problem in the dataset was conquered.

Sarcasm is another form of bullying; it always signifies the opposite meaning of what it states [75]. Due to this deceptive nature, sarcasm detection is still challenging for humans and machines. Many researchers work in this area and detect malicious content over the positive sentiment with high efficacy. Rendalkar et al. [76] proposed an emotion detection methodology for sarcasm detection on online social media. Libraries like WordNet and SentiWordNet are used to find textual similarity and emotion detection. The authors suggested using a combined methodology comprising different algorithms to identify whether the comment is sarcastic or not. Shrivastava et al. [77] proposed a method based on Google BERT (Bidirectional Encoder Representations from Transformers) to construct a sarcasm detection system. BERT is an open-source pre-trained ML framework that helps understand the meaning of the ambiguous nature of the text by analyzing the surrounding text. The main advantage of this model is that it can handle a large volume and variety of data efficiently and perform better than traditional approaches.

## 5.2 Identifying cyberbullying on multimodal data

Cyberbullying is not only limited to textual content; it may also spread through images or videos. OSM's users post memes (pictures or videos) to get the victim in trouble or harm their reputation. So it needs to be controlled quickly and effectively, particularly in this area. The research in this area is limited due to a lack of data availability or poor datasets. Cheng et al. [78] proposed the XBully framework for cyberbullying detection. The model initially looked for hotspots on a social media-based dataset and used the neighborhood and co-existence relations to develop a heterogeneous network. It was further divided into subnetworks, each consisting of 2 models. Nodes in each subnetwork were then standardized in a common latent space. Embeddings were chained, and ML models like the RF, linear SVM, and LR were trained with the XBully. It was concluded that Xbully outperformed Raw, DeepWalk, Node2Vec, GraRep, and Variant methods. Paul et al. [58] proposed a multimodal architecture based on an early identification approach, considering pictorial and textual entities in a post to detect cyberbullying. Comments under a particular post were processed sequentially to classify a post as non-bully or bully. Similarly, videos were processed frame by frame to develop an early detection solution. A unique text embedding was synthesized through the Transformer Encoder of Universal Sentence Encoder (TEUSE), and Principle Component Analysis (PCA) was used for dimensionality reduction. Finally, features from images and text were fused and fed into the Residual BLSTM Recurrent Neural Network and the model achieved an F-measure of 0.75. Wang et al. [57] also proposed a multimodal structure that considers a wholesome input consisting of videos, images, comments and time of the post to tackle real-time cyberbullying. This method outperformed the conventional text-based cyberbullying detection techniques, which could not satiate the all-around bullying data available on social media platforms. The model used BLSTM to extract features from a post Hierarchical Attention Network (HAN) to analyze comments and MLP for images and videos. Suryavanshi et al. [79] devised MultiOFF, a multimodal dataset to detect objectionable content consisting of memes based on 2016 U.S. Presidential polls. The dataset proved to be the forte of the research. Image and text entities were combined in an early fusion fashion and then compared with image-specific and text-specific outcomes. Pre-trained VGG16, a CNN-based model, was used to extract pictorial features and GloVe for word embedding. Comparisons were carried out between text classifiers (NB, LR, DNN, stacked LSTM, BLSTM, CNN), VGG16 image classifier, and multimodal classifiers (stacked LSTM + VGG16, BLSTM + VGG16, CNN on text + VGG16). The multimodal approach proved to be better than image-specific classifiers. Text classifiers were at par with multimodal classifiers and even better at times. The research concluded that weighing more on the textual features while combining them with pictorial features gives better results. Kumari et al. [53] proposed a combined portrayal of image and text rather than image-only and text-only classifiers. CNN with a single layer was trained with unified pictorial and textual features, giving better outcomes than a dual dimension representation. Each image was converted into a 3D matrix, and the words were vectorized using TF-IDF representation. The words represented in a single dimension were transformed into a 3D matrix to feed image and text data into CNN. The model predicted correctly for 74% of outcomes. It was observed that a single-layer convolution with a larger filter size outperforms a multiple-layer convolution with a smaller number of filters. This research also points to the need to weigh text and images for predictions. Yuvaraj et al. [80] employed a dual-engine model for cyberbullying detection. The model was unique as it takes psychological features apart from user comments and context. Furthermore, Artificial Neural Network was used to carry out classification and Deep Reinforcement learning to provide feedback after every loop. Hence, the model improved the results after every iteration. The ANN-DRL model illustrated an accuracy of 80.69%. Kumar et al. [81] defined a hybrid model that uses a capsule network (CapsNet) and dynamic routing to carry out textual bullying predictions and CNN for visual bullying predictions. Textual and visual entities were processed using this hybrid structure, and the late fusion decision layer was used for end prediction. Google Lens was used to segregate text pieces from images. However, the model had a limited scope in the real-time scenario where the data have high dimensionality and is multilingual. Karimvand et al. [82] proposed a multimodal DL technique involving a bi-directional gated recurrent unit (Bi-GRU) for analyzing textual data and 2-dimensional CNN (2CNN) for extracting meaningful features from images. A new dataset MPerIns was used to carry out training and testing, but its size is smaller than most of the datasets, which is a limitation. The textual analysis also comprised word embedding to convert text into numerical data. The deep fusion technique outperformed conventional ML models.

Sarcasm poses a significant challenge for the opinion mining system. Therefore, the multimodal sarcasm detection system has gained more attention. The main aim of this system is to understand the sentiment in images or videos. Sangwan et al. [83] defined the sarcasm detection model for multimodal datasets. The authors proposed an efficient method based on DL that uses both the info, i.e., textual and visual, for multimodal sarcasm detection. The main aim of this framework is the utilization of interdependencies of texts and images for the detection system. The proposed approach is based on recurrent neural networks that mainly use the interaction among the input modalities for the prediction. Results suggest that incorporating visual modalities has a key role in performance improvement. Yao et al. [84] proposed a multimodal, multi-interactive, and multi-hierarchical neural network. Twitter image, text in the image, and image caption are selected as inputs of this neural network as the brain's perception of sarcasm requires multiple modalities. A two-hierarchical structure is used, which leverages self-attention accompanied by attention pooling to integrate multimodal semantic information from different levels mimicking the brain's first- and second-order comprehension of sarcasm. Wu et al. [85] employed the incongruity-aware attention network (IWAN), which detects sarcasm by directing word-level incongruity between modalities via a scoring mechanism. This scoring mechanism could assign larger weights to words with incongruent modalities. Experimental results demonstrate the effectiveness of their proposed IWAN model on the MUStARD dataset and offer interpretability advantages.

Table 2. summarizes the related work communicated in various techniques performed on different datasets and collectively used their results to gain comparable knowledge. We also highlight each technique's advantages and limitations to suggest a direction for new research in this area. However, it may be noted that the numbers are not a clear sign of supremacy.

In 2011, Dinakar et al. [35] proposed the Ml classifier like NB, Rule-Based Jrip, Tree-Based J48, and SVM for textual cyberbullying detection. They experimented with their result based on the binary and multiclass classifiers and showed that binary class classifier performs better than multiclass classifier.

In 2011, Reynolds et al. [41] developed various classifiers on Formspring. me website simultaneously observed that DT and KN achieved 78.5% accuracy. Another study was proposed by Chen et al. [33] in 2012, who discussed the Lexical Syntactic Feature (LSF) new architecture with an SVM classifier. They had obtained the running time advantage with improved performance of the classifier.

In 2015 Mangaonkar et al. [26] developed the collaborative paradigm rather than the stand-alone paradigm to improve detection. The authors suggested that the collaborative approach performs well with running time advantages. In 2015, Wang et al. [60] had directly examined the relationship between emoticons and sentiment polarity. Emoticons directly express the sentiment expression, and negative sentiment expression always leads the cyberbullying. The authors performed the NB classifier to prove that the classifier with emoticons acquired better accuracy than those without emoticons.

In 2016, Chu et al. [50] employed a model for classifying offensive flag comments using DL-based models and achieved the highest *F*-score of 73%. They argued that using DL models gives better performance than non-DL models. Another author, Iwendi et al. [55], compared 4 DL models—BLSTM, GRU, LSTM, and RNN—to determine their effectiveness in cyberbullying detection. They found that BLSTM obtained the highest *F*-score of 88%. Another study was presented by Agrawal et al. [48] in 2018, who considered the three different social media platforms. They experimented with deep neural network-based models (CNN, LSTM, BLSTM, and BLSTM with attention). Initially, the authors used these models with the word embedding method. They noticed that BLSTM acquired the 91% *F*-score with attention and said that the word embedding method has not significantly affected cyberbullying detection. Later on, the authors used the transfer learning approach to gain knowledge from DL models on one dataset and improve the performance of the other. The authors found that BLSTM with

**Table 2** Summary of closely Related to our study

| Refs | Data type | Dataset | Classifier | Result in % | | | | Advantages | Observations |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Accuracy | Precision | Recall | *F*-measure | | |
| [35] | Textual Data | YouTube (4,500 Comments) | NB | 63 | NA | NA | NA | Binary classifiers perform better than multiclass classifiers | Sarcasm is frequently misclassified |
| | | | Rule-based Jrip | 63 | NA | NA | NA | | |
| | | | Tree-based J48 | 61 | NA | NA | NA | | |
| | | | SVM | 66.70 | NA | NA | NA | | |
| [41] | | Formspring.me (18,554 user's information) | DT(J48) | 78.5 | NA | NA | NA | Perform well on a small data set | Classify lots of false-negative if training data is less than 10% true positive value |
| | | | Rule-based Jrip | 77.3 | NA | NA | NA | | |
| | | | KNN | 78.5 | NA | NA | NA | | |
| | | | SVM | 67.2 | NA | NA | NA | | |
| [33] | | YouTube (2,175,474 comments) | LSF SVM | NA | 98.24 | 94.34 | 96.28 | The running time of LSF SVM is less than the existing model | Misclassify the sentence due to nouns and pronouns in parts because of the input's typographical type of errors |
| [26] | | Twitter (1340 tweets) | NB (bi-gram) | 76 | 80 | 67 | 72.92 | Model perform well with running time advantage | SVM fails when the dataset is unbalanced |
| | | | LR (bi-gram) | 73 | 71 | 85 | 77.37 | | |
| | | | SVM (bi-gram) | 65 | 64 | 88.5 | 74.28 | | |
| [60] | | Twitter (500 tweets) | NB with emoticon | 78 | 84 | 87 | 86 | Directly examine the relationship between emoticons and sentiment expression | Only applicable for a small dataset |
| | | | NB without emoticon | 61 | 54 | 56 | 54 | | |
| [50] | | Wikipedia (150 K comments) | RNN-LSTM and word embedding | 93 | NA | NA | 70 | Deep learning-based models Perform better than traditional approaches | CNN-character embedding requires more training time as compared with the two others |
| | | | CNN-word embedding | 93 | NA | NA | 70 | | |
| | | | CNN-character embedding | 94 | NA | NA | 73 | | |
| [55] | | Kaggle | BLSTM | 82.18 | 86 | 91 | 88 | BLSTM performs better than empirical LSTM | The remaining models do not achieve acceptable results |
| | | | GRU | 81.46 | 86 | 89 | 87 | | |
| | | | LSTM | 80.86 | 85 | 90 | 87 | | |
| | | | RNN | 81.01 | 85 | 90 | 87 | | |

**Table 2** (continued)

| Refs | Data type | Dataset | Classifier | Result in % | | | | Advantages | Observations |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Accuracy | Precision | Recall | *F*-measure | | |
| [48] | | Formspring.me (12 k posts), Twitter (16 k posts), Wikipedia (100 k posts) | CNN | NA | 93 | 90 | 91 | This model performs better than the empirical model | LSTM is least efficient than other models |
| | | | LSTM | NA | 91 | 85 | 88 | | |
| | | | BLSTM | NA | 91 | 81 | 86 | | |
| | | | $\text{BLSTM}_{\text{with attention}}$ | NA | 90 | 91 | 91 | | |
| [86] | | Twitter | GRU | 95.25 | NA | NA | 64.8 | The ULMFiT model significantly outperformed the other Architectures | Aggressive fine-tuning may be suffering from catastrophic forgetting |
| | | | CNN | 94 | NA | NA | 64 | | |
| | | | ULMFiT | 96.7 | NA | NA | 96.67 | | |
| [22] | Imbalanced Textual Data | Kaggle (4000 comments) | LR | 83.76 | 64.4 | 61.47 | 62.90 | Easily classify the unlabelled data set | It Cannot classify the sarcastic comments |
| | | | SVM | 77.65 | 70.29 | 58.29 | 61.72 | | |
| [73] | | Twitter (10,606 tweets) | NB | NA | 76.30 | 77.4 | 76.8 | Model use with SMOTE technique to classify imbalanced data with high accuracy | The performance of the NB classifier is not satisfactory |
| | | | libSVM | NA | 82 | 83.1 | 78.6 | | |
| | | | RF | NA | 94.10 | 93.9 | 93.6 | | |
| | | | KNN | NA | 87 | 87.3 | 87.1 | | |
| [24] | | Twitter (16,914 tweets) | TF-IDF + SVM | NA | 78.7 | 62.9 | 65.6 | The Char-CNNS model easily classifies the spelling errors, noisy, short, and unstructured data, and the dataset's imbalanced class problem | The model is only suitable for textual datasets |
| | | | Word n-gram + LR | NA | 72.23 | 54.5 | 58.8 | | |
| | | | Char n-gram + LR | NA | 72.8 | 68.9 | 67.3 | | |
| | | | Word-CNN | NA | 75.2 | 68.4 | 72.5 | | |
| | | | Char-CNNS | NA | 81 | 70.5 | 74.2 | | |
| [47] | Multilingual Textual and Visual Data | Facebook, Twitter (35,273 posts) | NB | NA | 36.5 | 30.4 | 33.2 | SVM performs better among all classifiers in terms of precision | The performance of the classifiers is not satisfactory |
| | | | SVM | NA | 81.5 | 27 | 40.5 | | |
| [59] | | Twitter (508 tweets) | Multinomial NB | 78.94 | 94.23 | 78.95 | 85.91 | Easily classify the multilingual text (Marathi) | Only applicable for small data sets |

**Table 2** (continued)

| Refs | Data type | Dataset | Classifier | Result in % | | | | Advantages | Observations |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Accuracy | Precision | Recall | *F*-measure | | |
| | | | LR | 82.36 | 93.23 | 82.36 | 89.54 | | |
| | | | Stochastic Gradient Descent | 81.57 | 94.33 | 81.58 | 87.49 | | |
| [25] | | Twitter (2,349,052 tweets) | Linear SVM | 48.8 | NA | NA | NA | Most classifiers predict bullying with multiple textual features accurately | The performance of linear SVM is not satisfactory |
| | | | LR | 93.4 | 93.4 | 93.6 | 93.5 | | |
| | | | DT | 90.2 | 88.2 | 93.2 | 90.6 | | |
| | | | RF | 91.9 | 91.8 | 92.5 | 92.2 | | |
| | | | Gradient Boosting regression tree | 93.3 | 92.4 | 94.6 | 93.5 | | |
| | | | Multilayer Perception | 90.9 | 90.9 | 91.3 | 91.1 | | |
| [87] | | | $^{\text{softAtt}}$ BLSTM | 92.71 | 89.49 | 90.67 | 89.05 | The proposed model classified sarcasm in code mix language effectively | Identifying a sense of online context in Hindi or English language depends on the part of the speech tagger and the language identifier, which the authors have not met |
| | | | BLSTM | 85.03 | 78.77 | 81.27 | 79.5 | | |
| | | | LSTM | 81.75 | 74.91 | 78.28 | 76.45 | | |
| [58] | Video data | Vine | BiLSTM | NA | 50 | 45 | 47 | Models perform well on multimodal dataset | It misclassifies the sarcasm comments or posts |
| | | | Recurrent-CNN | NA | 57 | 67 | 62 | | |
| | | | BiLSTM-RecurrentCNN | NA | 72 | 55 | 69 | | |
| | | | ResBiLSTM-RCNN | NA | 75 | 75 | 75 | | |
| [57] | | Vine | MMCD | 83.8 | NA | NA | 84.1 | The proposed model efficiently works on the imbalanced dataset | The proposed model is not performing well when the learning rate is exceptionally high or low |
| [88] | Image and Text Data | Instagram (2188 posts) | CNN + TF-IDF | 77 | 79 | 75 | 76 | The proposed model significantly outperformed textual and contextual data | The NN combiner has performed well for all the evaluation metrics other than Recall |
| | | | CNN + Word2Vec | 78 | 79 | 76 | 77 | | |
| | | | LSTM | 83 | 85 | 84 | 84 | | |
| | | | NN combiner | 86 | 87 | 83 | 85 | | |

**Table 2** (continued)

| Refs | Data type | Dataset | Classifier | Result in % | | | | Advantages | Observations |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Accuracy | Precision | Recall | *F*-measure | | |
| | | | Boosting combiner | 84 | 83 | 85 | 84 | | |
| [81] | | YouTube, Instagram, Twitter (10,000 comments) | KNN | 65.8 | 71.3 | 71.8 | NA | The proposed architecture handles the textual, visual, and infographic modalities together | The proposed model cannot handle the skewed, heterogeneous, and imbalanced dataset |
| | | | NB | 64.4 | 67 | 69.2 | NA | | |
| | | | SVM | 73.2 | 76.86 | 79.17 | NA | | |
| | | | ConvNet | 97.05 | 98.6 | 95.08 | NA | | |
| [89] | | Facebook (41,350 posts) | SVM | 88.39 | NA | NA | NA | The proposed model easily handled the noisy data | The Gaussian NB performed the worst among all classifiers |
| | | | Ada Boost | 90.61 | NA | NA | NA | | |
| | | | RF | 93.11 | NA | NA | NA | | |
| | | | Multilayer Perceptron | 92.06 | NA | NA | NA | | |
| | | | Gaussian NB | 73.66 | NA | NA | NA | | |
| [90] | | News website (10,000 articles) | ELA + CNN | 44.20 | NA | NA | 51.86 | The ViLBERT model outperformed the Visio linguistic (textual and visual) modalities | The Image forgery detection-based model performed worst |
| | | | ViLBERT | 93.80 | NA | NA | 92.16 | | |

attention achieved the highest *F*-score of 94% for two out of 3 datasets. The authors [86] presented another deep learning approach in 2019: GRU, CNN, and Universal Language Model Fine Tuning(ULMFiT*)*. The ULMFiT model uses transfer learning and fine-tuning techniques and achieved an average *F*-score of 96.67%.

In 2015, Chavan et al. [22] involved two new features; capturing pronouns and using the skip grams for aggressive comment detection on OSM. They improved the model's performance and gained a precision of 70% and an area under the curve (AUC) of 86.92%, but their models do not handle the imbalanced dataset. In 2016, another methodology was developed by Al-garadi et al. [73] for handling the imbalanced dataset using the SMOTE techniques; and they also included new features such as network, activity, user, Twitter textual content with the improved result via applying NB, libSVM, RF, and KNN. It was observed that the authors achieved an approximate 94% precision and AUC of 94.3%. In 2020, the author Lu et al. [24] dealt with a shortcut, spelling error, and unstructured data problem with the applicability of ML and DL models. They used the character as the smallest learning unit in models to overcome this problem and stated that short, noisy, unstructured data with a spelling error are also detectable. Still, their model did not perform well.

In 2017, Haider et al. [47] deployed a model for aggressive multilingual content. The authors proved that detecting bullying content in languages other than English is also possible. Another study was proposed by Pawar et al. [59] in 2019, who used the synthetic data generation technique to solve the class imbalance problem in the multilingual cyberbullying detection process. Synthetic data generation increases the classifier's performance and gains an accuracy of 82.36%. The other author, Zhang et al. [25], considered the multiple textual features and improved the accuracy by over 90% on multilingual text. In 2020, Jain et al. [87] presented a multimodal sarcasm detection for code-mix (Hindi and English). They had used the BLSTM with a softmax attention layer and feature-rich CNN model. These two models use the Hindi, English, and Pragmatic feature vectors. They used the Hindi-SentiWordNet to generate the senti-Hindi feature vector, BLSTM with attention to generate the English context vector, and auxiliary pragmatic feature vector to count the total pragmatic marker in tweets. Their model acquired a classification accuracy of 92.71% and an *F*-score of 89.05% on code switch tweets.

In 2020, Paul et al. [58] designed the DL-based model for the multimodal dataset (Vine). The authors used the TEUSE for text embedding on BLSTM, GRU, LSTM, RNN, BiLSTM, and Recurrent-CNN and achieved an *F*-score of 75%. Another study which was proposed by Wang et al. [57], had also employed the Multi-Modal Cyberbullying Detection (MMCD) framework, and they used two datasets: Instagram (image and video) and Vine (short videos). The authors used the HAN and MLP techniques with BLSTM to improve the efficacy. They achieved improved accuracy and F-score on an average of 83% and 84% simultaneously on the Vine dataset.

In 2020, Rezvani et al. [88] proposed a cyberbullying detection model for metadata (image and text). In this direction, the author first extracted the features from images, the Image's metadata information, and generated comments and text around the images. After that, they contextualized the extracted features using a crowdsourced feedback loop. Finally, they combined all the features using a neural network model to identify the most important features. The author used the CNN and LSTM models and achieved an efficiency of 86%. Another study was proposed by Kumar et al. [81] in 2021; the authors designed a hybrid (CapsNet-ConvNet) model for textual, visual, and infographic (Text embedded along with image). They gained improved accuracy with 97.05% accuracy on infographic data as well.

In 2018, Das et al. [89] employed a model for sarcasm detection on image data. The authors used the Facebook Graph API for data collection and considered the posted texts, images, comments on the post, and count of user interaction on that post. They used the BOW for feature extraction and IG for feature selection with Supervised learning classifiers like SVM, Ada Boost, RF, Multilayer Perceptron (MLP), and Gaussian NB. They have achieved the 93% highest accuracy. In 2020, Li et al. [90] constructed the Visio linguistic model for satire detection. They collected the data (Images and Headlines) from news websites, used the Vision and Language BERT (ViLBERT) model with image forensics technique, and improved accuracy by 93.8%.

## 6 Challenges and issues

This section discusses the issues and challenges in designing a cyberbullying prediction system. We also address these challenges as follows.

### 6.1 Cyberbullying identification

Identifying a cyberbullying act is a big challenge for a cyberbullying prediction system [91]. Some social media users comment or post in sarcastic ways, such as *"oh darling, go buy a personality,"* making it difficult to classify as offensive content without analyzing a logical factor. However, this post or comment represents a negative sentiment [92]. Expressions of sarcastic are challenging to diagnose in sentiment analysis, and automating the process further complicates the process. Since sarcasm signifies the opposite of what it states, obtaining a clear context from sarcastic expressions is the main hurdle that needs to be tackled.

Kumar et al. [93] compared some ML and DL models using Twitter SemEval 2015 task 11 and Reddit data and came up with the result that BLSTM outperforms other models obtaining an accuracy of 86.32% and 82.91% on the datasets, respectively. We suggest dealing with the text's logic and semantics; building the appropriate dictionary with new keywords and using word2vec to learn word associations is essential. W2C (word2vec) performs better for a larger vocabulary and is more tweakable in our case. It allows tuning our trained word vector to a specific application.

### 6.2 Slang identification

Slang is a language regarded as very informal and used within various subcultures. For example, the word *dope* conventionally means a recreational drug consumed illegally but can also mean "very good" based on the context. This challenge aims to find the proper usage of slang in a sentence and identify its exact location. Many new slang terms are deliberately misspelled or missing characters. Quick changes in language by youngsters can impact keywords used as a feature in cyberbullying detection. As recorded in a paper by Pei et al. [94], we need to work with two kinds of slang; newly extended senses and newly created words. As a result, the language corpus should regularly evolve dynamically, and ML algorithms should be retrained. The research also proposed the BLSTM model that pinpointed the position of the slang, which had better results than previously proposed models. We can use Glove Twitter word embedding for this, but Wilson et al. [95] made a corpus from Urban Dictionary (UD) with more modern slang for our reference. GRU might perform better here. GRU has two gates (reset and update), unlike LSTM. It does not have a cell state and uses a hidden state instead. It has to perform fewer operations so it can be trained much faster.

### 6.3 Translational ambiguity

Further, cyberbullying is not limited to English only, but it can also happen in other languages like Hindi, Bengali, Arabic, Spanish, Hinglish, etc. Due to language diversity on social media, text classification can be challenging in multilingual input (for example, Hinglish) as literal translation can be ambiguous. We can translate the input into the language of the dataset. Still, this approach has some shortcomings (Google translates the Hinglish phrase—"Yeh ladki ekdum chaalu hai" to "This girl is on the move") as phrases are not precisely translated well by various translation tools. Kumar et al. [56] proposed using a language-dependent model with a dataset of different modern languages. Google trans-literation toolkit is used to convert the multilingual phrase into its original language (Hinglish—"Yeh ladki ekdum chaalu hai" to Hindi-"यह लड़की एकदम चालु है"). It allows us to interpret the phrase correctly.

### 6.4 Feature extraction and selection

Most of the techniques used to carry out cyberbullying detection have been improvised by increasing the available data features. However, this poses a problem of making feature extraction exceedingly onerous apart from increasing the computational time. Al-Ajlan et al. [49] came up with an approach that substitutes feature extraction with word vectors. Glo-Ve was the word embedding technique used, metaheuristic optimization to decide the optimum parameters, then fed into CNN. CNN has a long-term dependency, so we will use the Gated Recurrent Unit (GRU) model. It mitigates the vanishing/exploding gradient and can remember information for more extended periods. It might be performing better than using CNN.

Feature selection is essential in removing irrelevant features or reducing dimensionality to increase learning accuracy [96]. An important factor is mainly selected, and this choice of feature selection depends on many characteristics of the dataset like data types, data size, noise, etc. [97]. Many techniques are available for feature selection. Each technique's main aim is to choose a subset of features to improve the prediction rate's effectiveness or reduce the structure's size without minimizing the prediction rate's effectiveness [98]. However, sometimes the most essential selected features may not be able to help a model learn. For instance, some offenders might not have used commonly traditional bullying words. It requires the model to dynamically adapt the new terms and acronyms for cyberbullying. If such types of words and acronyms are not considered a dependent variable for feature selection, it may affect prediction accuracy. Some other factors, such as age, gender, occupation, and religion, provided by the social media user are included as features to improve cyberbullying classifier performance accuracy. Dadvar et al. [99] included gender and age to enhance ML classifier performance. Here, the issue is that most malicious users would not reveal exact details or behave like anonymous users.

### 6.5 Offensive image detection

Cyberbullying can also occur in the form of offensive memes and profane images. Identifying such memes and indecent images is also a very challenging task. Suryawanshi et al. [79] suggested using CNN and BiLSTM models for detecting appropriate and inappropriate images. Tumblr (a microblogging website) dataset can be used in training the model. The authors used the MultiOFF dataset, which has memes from the 2016 US presidential elections from various websites such as Facebook and Reddit. We suggest using the GRU model, which might give better results than CNN and BiLSTM. GRU can be trained faster than LSTM and performs better than LSTM for smaller datasets. It also does not have the problem of vanishing gradients that CNN has.

The above-discussed challenges and possible solutions are directly related to input data like textual or visual. Some issues are related to textual data, some are related to visual data, and some are related to both. So it is a very challenging task to design a standard cyberbullying detection model for multimodality input while considering these challenges. There is no well-defined solution covering all the above challenges of cyberbullying detection through a single framework. Therefore, we plan to build a comprehensive architecture considering all difficulties discussed in our subsequent research work.

Following are the findings of this review work that can address the challenges mentioned above and be incorporated to develop a novel architecture for effective cyberbullying detection.

The first phase is data collection. The main challenge of this phase is the identification of cyberbullying content. Sarcastic, slang, and code mix language are the main issue in cyberbullying identification. So we need to require specialized tools and techniques to collect, analyze and build an appropriate dictionary with new keywords. The language corpus should be dynamically updated regularly. We suggest using an Urban dictionary for bullying identification because it has more modern toxic keywords for slang words. We should create a corpus for code mix language and regularly update it with new profane words.

The second phase is data preprocessing. We clean the data, divide data into tokens, use the root word, convert the lower to upper, and vice versa. We need to follow a standard process for every technique. We suggest using lemmatization instead of stemming because lemmatization analyses a broader aspect of what a word means in a sentence by just recognizing its root word.

Feature extraction and selection is the third and most important phase of a system. In the feature extraction techniques, we select the useful feature from row data while filtering the essential feature from the selected feature in feature selection. Therefore, we suggest using Word2vec to learn word association for the ML-based model and Glo-Ve for the DL-based model for feature extraction. Consider the essential feature meta-information for feature selection: age, gender, occupation, religion, area, number of followers, etc.

The last phase is the classification algorithm. The choice of classification algorithm depends on dataset characteristics such as datatype, size of data, noise, etc. Although ML-based classification and DL-based classification both are efficiently working on textual data. However, the DL-based classification is most suited for visual data. Therefore, we suggest using RF, BLSTM, and GRU for textual and GAIN and GRU for visual modalities.

## 7 Conclusion

Our survey thoroughly reviewed the existing ML and DL classification techniques to predict and detect offensive online content, which is essential to protect genuine online users. In this paper, the most common forms of cyberbullying have been elaborated on in-depth. It can always help one identify and distinguish acts of cyberbullying based on a person's activities, and it also aids a victim in acknowledging if they are being bullied. We briefly described all the practical steps required to build a cyberbullying detection system. Machine learning and Deep Learning were also summarized to detect and predict offensive content on social media sites. We also investigated the dataset extraction methods from social media and explained some supporting tools used for data preprocessing, data labeling, and classification. Further, we divided our study into two sections based on the data type. We comprehensively analyzed the various factors in each section to construct an effective classifier in this area. We specifically also tried to address issues and challenges in this area and mentioned the findings for building a cyberbullying detection model. Thus, this review will provide crucial details and the right direction to novice researchers looking forward to working in this field.

**Funding** Any agency does not fund this study.

**Data availability statement** Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

## Declarations

**Conflict of interest** All the authors of this paper declare that he/she has no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

1. Cénat, J.M., Hébert, M., Blais, M., Lavoie, F., Guerrier, M., Derivois, D.: Cyberbullying, psychological distress and self-esteem among youth in Quebec schools. J. Affect. Disord. **169**, 7–9 (2014)
2. Al-Garadi, M.A., Hussain, M.R., Khan, N., Murtaza, G., Nweke, H.F., Ali, I., Gani, A.: Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. IEEE Access **7**, 70701–70718 (2019)

3. Zhao, R., Mao, K.: Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. IEEE Trans. Affect. Comput. **8**(3), 328–339 (2016)
4. "UNICEF POLL:More than a third of young people in 30 countries report being a victim of online bullying",2019,[online] Available: https://www.unicef.org/press-releases/unicef-poll-more-third-young-people-30-countries-report-being-victim-online-bullying
5. "Cyberbullying Statistics", 2019,[online] Available: https://enough.org/stats_cyberbullying
6. Summary of Our Cyberbullying Research. 2007–2019,[online]https://cyberbullying.org/summary-of-our-cyberbullying-research
7. Sargar, B., Kattimani, R.: Emerging trends and issues in social sciences (2020)
8. Sampasa-Kanyinga, H., Roumeliotis, P., Xu, H.: Associations between cyberbullying and school bullying victimization and suicidal ideation, plans and attempts among Canadian schoolchildren. PLoS ONE **9**(7), e102145 (2014)
9. Hinduja, S., Patchin, J.W.: Bullying, cyberbullying, and suicide. Arch. Suicide Res. **14**(3), 206–221 (2010)
10. Hosseinmardi, H., Li, S., Yang, Z., Lv, Q., Rafiq, R.I., Han, R., & Mishra, S.: A comparison of common users across instagram and ask. fm to better understand cyberbullying. In 2014 IEEE Fourth International Conference on Big Data and Cloud Computing (pp. 355–362). IEEE (2014)
11. Maher, D.: Cyberbullying: an ethnographic case study of one Australian upper primary school class. Youth Stud. Austr. **27**(4), 50 (2008)
12. Willard, N.E.: Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Press, Delhi (2007)
13. Patchin, J.W., Hinduja, S.: Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth Violence Juvenile Justice **4**(2), 148–169 (2006)
14. Bayzick, J., Kontostathis, A., Edwards, L.: Detecting the presence of cyberbullying using computer software (2011)
15. Bishop, J.: Tackling Internet abuse in Great Britain: Towards a framework for classifying severities of'flame trolling'. In: Proceedings of the International Conference on Security and Management (SAM) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) (2012)
16. Bishop, J.: Representations of “trolls” in mass media communication: a review of media-texts and moral panics relating to “internet trolling.” Int. J. Web Based Commun. **10**(1), 7–24 (2014)
17. Bhat, A.: An analysis of crime data under apache pig on big data. In: 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 330–335). IEEE (2019)
18. "Cyberbullying Statistics, Facts and Trend. [online] https://firstsiteguide.com/cyberbullying-stats/ (2020)
19. González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., Moreno, Y.: Assessing the bias in samples of large online networks. Social Netw. **38**, 16–27 (2014)
20. Akhter, S.: Social media bullying detection using machine learning on Bangla text. In: 2018 10th International Conference on Electrical and Computer Engineering (ICECE) (pp. 385–388). IEEE (2018)
21. Monika, A., Bhat, A.: Automatic Twitter crime prediction using hybrid wavelet convolutional neural network with world cup optimization. Int. J. Pattern Recognit. Artif. Intell. *36*(05), 2259005 (2022)
22. Chavan, V.S., & Shylaja, S.S.: Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2354–2358). IEEE (2015)
23. Eshan, S.C., & Hasan, M.S.: An application of machine learning to detect abusive bengali text. In: 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1–6). IEEE (2017). https://doi.org/10.1109/ICCITECHN.2017.8281787
24. Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., Choo, K.K.R.: Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. Concurr. Comput. Pract. Exp. **32**(23), e5627 (2020)
25. Zhang, J., Otomo, T., Li, L., Nakajima, S.: Cyberbullying detection on twitter using multiple textual features. In: 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST) (pp. 1–6). IEEE (2019)
26. Mangaonkar, A., Hayrapetian, A., & Raje, R.: Collaborative detection of cyberbullying behavior in Twitter data. In 2015 IEEE International Conference on Electro/Information Technology (EIT) (pp. 611–616). IEEE (2015)
27. Xue, B., Fu, C., Shaobin, Z.: A study on sentiment computing and classification of sina weibo with word2vec. In: 2014 IEEE International Congress on Big Data (pp. 358–363). IEEE (2014)
28. Bozyiğit, A., Utku, S., Nasibov, E.: Cyberbullying detection: utilizing social media features. Expert Syst. Appl. **179**, 115001 (2021)
29. Isaac, A., Kumar, R., Bhat, A.: Hate speech detection using machine learning techniques. In: Sustainable Advanced Computing (pp. 125–135). Springer, Singapore (2022)
30. Nahar, V., Al-Maskari, S., Li, X., Pang, C.: Semi-supervised learning for cyberbullying detection in social networks. In Australasian Database Conference (pp. 160–171). Springer, Cham (2014).
31. Nalini, K., Sheela, L.J.: Classification using latent dirichlet allocation with Naive Bayes classifier to detect cyber bullying in Twitter. Indian J. Sci. Technol. **9**(28), 1–5 (2016)
32. Balakrishnan, V., Khan, S., Arabnia, H.R.: Improving cyberbullying detection using Twitter users' psychological features and machine learning. Comput. Secur. 101710 (2020)
33. Chen, Y., Zhou, Y., Zhu, S., & Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing (pp. 71–80). IEEE (2012)
34. Galán-García, P., Puerta, J.G.D.L., Gómez, C.L., Santos, I., Bringas, P.G.: Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. Logic J. IGPL **24**(1), 42–53 (2016)
35. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
36. Sanchez, H., Kumar, S.: Twitter bullying detection. Ser. NSDI **12**(2011), 15 (2011)
37. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European Conference on Machine Learning (pp. 137–142). Springer, Berlin, Heidelberg (1998)
38. Andriansyah, M., Akbar, A., Ahwan, A., Gilani, N. A., Nugraha, A. R., Sari, R. N., & Senjaya, R. (2017, November). Cyberbullying comment classification on Indonesian Selebgram using support vector machine method. In *2017 Second International Conference on Informatics and Computing (ICIC)* (pp. 1–5). IEEE.

39. Ahmed, M., Goel, M., Kumar, R., & Bhat, A.: Sentiment analysis on Twitter using ordinal regression. In: 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) (pp. 1–4). IEEE (2021)
40. León-Paredes, G.A., Palomeque-León, W.F., Gallegos-Segovia, P.L., Vintimilla-Tapia, P.E., Bravo-Torres, J.F., Barbosa-Santillán, L.I., Paredes-Pinos, M.M.: Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language. In: 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON) (pp. 1–7). IEEE (2019)
41. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. In: 2011 10th International Conference on Machine Learning and Applications and Workshops (Vol. 2, pp. 241–244). IEEE (2011).
42. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (pp. 672–680).
43. García-Recuero, Á.: Discouraging abusive behavior in privacy-preserving online social networking applications. In: Proceedings of the 25th International Conference Companion on World Wide Web (pp. 305–309). International World Wide Web Conferences Steering Committee (2016)
44. Fazil, M., Abulaish, M.: A hybrid approach for detecting automated spammers in twitter. IEEE Trans. Inf. Forensics Secur. **13**(11), 2707–2719 (2018)
45. Soucy, P., Mineau, G.W.: A simple KNN algorithm for text categorization. In: Proceedings 2001 IEEE International Conference on Data Mining (pp. 647–648). IEEE (2001)
46. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Trans. Evol. Comput. **1**(1), 67–82 (1997)
47. Haidar, B., Chamoun, M., & Serhrouchni, A.: Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content. In: 2017 1st Cyber Security in Networking Conference (CSNet) (pp. 1–8). IEEE (2017)
48. Agrawal, S., Awekar, A.: Deep learning for detecting cyberbullying across multiple social media platforms. In: European Conference on Information Retrieval (pp. 141–153). Springer, Cham (2018)
49. Al-Ajlan, M.A., Ykhlef, M.: Optimized Twitter cyberbullying detection based on deep learning. In: 2018 21st Saudi Computer Society National Computer Conference (NCC) (pp. 1–5). IEEE (2018). https://doi.org/10.1109/NCG.2018.8593146
50. Chu, T., Jue, K., Wang, M.: Comment abuse classification with deep learning. *Von* https://web.stanford.edu/class/cs224n/reports/2762092.pdfabgerufen. *(2016)*
51. Mahlangu, T., Tu, C.: Deep learning cyberbullying detection using stacked embbedings approach. In: 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI) (pp. 45–49). IEEE (2019)
52. Haidar, B., Chamoun, M., Serhrouchni, A.: Arabic cyberbullying detection: using deep learning. In: 2018 7th International Conference on Computer and Communication Engineering (ICCCE) (pp. 284–289). IEEE (2018)
53. Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: Towards Cyberbullying-free social media in smart cities: a unified multimodal approach. Soft. Comput. **24**(15), 11059–11070 (2020)
54. Fang, Y., Yang, S., Zhao, B., Huang, C.: Cyberbullying detection in social networks using Bi-gru with self-attention mechanism. Information **12**(4), 171 (2021)
55. Iwendi, C., Srivastava, G., Khan, S., Maddikunta, P.K.R.: Cyberbullying detection solutions based on deep learning architectures Multimed. Syst. 1–14 (2020)
56. Kumar, A., & Sachdeva, N.: Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. Multimed Syst. 1–15 (2020)
57. Wang, K., Xiong, Q., Wu, C., Gao, M., & Yu, Y.: Multimodal cyberbullying detection on social networks. In: 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). IEEE (2020)
58. Paul, S., Saha, S., Hasanuzzaman, M.: Identification of cyberbullying: a deep learning based multimodal approach. Multimed. Tools Appl., 1–20 (2020). https://doi.org/10.1007/s11042-020-09631-w
59. Pawar, R., & Raje, R.R.: Multilingual cyberbullying detection system. In: 2019 IEEE International Conference on Electro Information Technology (EIT) (pp. 040–044). IEEE (2019)
60. Wang, H., Castanon, J.A.: Sentiment expression via emoticons on social media. In: 2015 IEEE International Conference on Big Data (Big Data) (pp. 2404–2408). IEEE (2015)
61. Singh, V.K., Huang, Q., Atrey, P.K.: Cyberbullying detection using probabilistic socio-textual information fusion. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 884–887). IEEE (2016). https://doi.org/10.1109/ASONAM.2016.7752342
62. Grasso, G., Furche, T., Schallhart, C.: Effective web scraping with oxpath. In: Proceedings of the 22nd International Conference on World Wide Web (pp. 23–26) (2013). https://doi.org/10.1145/2487788.2487796
63. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Hoste, V.: Automatic detection and prevention of cyberbullying. In: International Conference on Human and Social Analytics (HUSO 2015) (pp. 13–18). IARIA (2015)
64. Haidar, B., Chamoun, M., Serhrouchni, A.: Arabic cyberbullying detection: enhancing performance by using ensemble machine learning. In: 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) (pp. 323–327). IEEE (2019)
65. Gupta, P., Kaushik, B.: Suicidal tendency on social media: a case study. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 273–276). IEEE (2019)
66. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: an online lexical database. Int. J. Lexicogr. **3**(4), 235–244 (1990)
67. AlMaayah, M., Sawalha, M., Abushariah, M.A.: Towards an automatic extraction of synonyms for Quranic Arabic WordNet. Int. J. Speech Technol. **19**(2), 177–189 (2016)
68. Fryling, M., Cotler, J.L., Rivituso, J., Mathews, L., Pratico, S.: Cyberbullying or normal game play? Impact of age, gender, and experience on cyberbullying in multiplayer online gaming environments: perceptions from one gaming forum. J. Inf. Syst. Appl. Res. **8**(1), 4 (2015)
69. Foong, Y.J., Oussalah, M.: Cyberbullying system detection and analysis. In: 2017 European Intelligence and Security Informatics Conference (EISIC) (pp. 40–46). IEEE (2017)
70. Kontostathis, A., Reynolds, K., Garron, A., Edwards, L.: Detecting cyberbullying: query terms and techniques. In: Proceedings of the 5th Annual acm Web Science Conference (pp. 195–204) (2013)
71. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. Proc. Content Anal. WEB **2**, 1–7 (2009)
72. Bosse, T., & Stam, S.: A normative agent system to prevent cyberbullying. In: Proceedings of the 2011 IEEE/WIC/ACM

International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 02 (pp. 425–430). IEEE Computer Society (2011)

73. Al-garadi, M.A., Varathan, K.D., Ravana, S.D.: Cybercrime detection in online communications: the experimental case of cyberbullying detection in the Twitter network. Comput. Hum. Behav. **63**, 433–443 (2016)
74. Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J.P., Kowalski, R., & Dillon, E.: Cyberbullying detection with a pronunciation based convolutional neural network. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 740–745). IEEE (2016)
75. Kumar, R., & Bhat, A.: An analysis on sarcasm detection over twitter during COVID-19. In: 2021 2nd International Conference for Emerging Technology (INCET) (pp. 1–6). IEEE (2021)
76. Rendalkar, S., & Chandankhede, C.: Sarcasm detection of online comments using emotion detection. In: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1244–1249). IEEE (2018)
77. Shrivastava, M., Kumar, S.: A pragmatic and intelligent model for sarcasm detection in social media text. Technol. Soc. **64**, 101489 (2021). https://doi.org/10.1016/j.techsoc.2020.101489
78. Cheng, L., Li, J., Silva, Y.N., Hall, D.L., Liu, H.: Xbully: cyberbullying detection within a multimodal context. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 339–347) (2019)
79. Suryawanshi, S., Chakravarthi, B.R., Arcan, M., Buitelaar, P.: Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (pp. 32–41) (2020).
80. Yuvaraj, N., Srihari, K., Dhiman, G., Somasundaram, K., Sharma, A., Rajeskannan, S., Masud, M.: Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking. Math. Probl. Eng. 2021 (2021)
81. Kumar, A., Sachdeva, N. (2021). Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network. Multimed. Syst. 1–10.
82. Karimvand, A.N., Chegeni, R.S., Basiri, M.E., Nemati, S.: Sentiment analysis of persian instagram post: a multimodal deep learning approach. In: 2021 7th International Conference on Web Research (ICWR) (pp. 137–141). IEEE (2021)
83. Sangwan, S., Akhtar, M.S., Behera, P., Ekbal, A.: I didn't mean what I wrote! Exploring Multimodality for Sarcasm Detection. In: 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). IEEE (2020)
84. Yao, F., Sun, X., Yu, H., Zhang, W., Liang, W., Fu, K.: Mimicking the brain's cognition of sarcasm from multidisciplines for Twitter sarcasm detection. IEEE Trans. Neural Netw. Learn. Syst. (2021)
85. Wu, Y., Zhao, Y., Lu, X., Qin, B., Wu, Y., Sheng, J., Li, J.: Modeling incongruity between modalities for multimodal sarcasm detection. IEEE MultiMed. (2021)
86. Amrutha, B.R., Bindu, K.R.: Detecting hate speech in tweets using different deep neural network architectures. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 923–926). IEEE (2019)
87. Jain, D., Kumar, A., Garg, G.: Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. Appl. Soft Comput. **91**, 106198 (2020)
88. Rezvani, N., Beheshti, A., Tabebordbar, A.: Linking textual and contextual features for intelligent cyberbullying detection in social media. In: Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia (pp. 3–10) (2020). https://doi.org/10.1145/3428690.3429171
89. Das, D., Clark, A.J.: Sarcasm detection on facebook: a supervised learning approach. In: Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct (pp. 1–5) (2018)
90. Li, L., Levi, O., Hosseini, P., Broniatowski, D.A.: A multimodal method for satire detection using textual and visual cues. arXiv:2010.06671. (2020). https://doi.org/10.48550/arXiv.2010.06671
91. El Asam, A., Samara, M.: Cyberbullying and the law: a review of psychological and legal challenges. Comput. Hum. Behav. **65**, 127–141 (2016). https://doi.org/10.1016/j.chb.2016.08.012
92. Salawu, S., He, Y., Lumsden, J.: Approaches to automated detection of cyberbullying: a survey. IEEE Trans. Affect. Comput. (2017)
93. Kumar, A., Garg, G.: Empirical study of shallow and deep learning models for sarcasm detection using context in benchmark datasets. J. Ambient Intell. Human. Comput. 1–16 (2019)
94. Pei, Z., Sun, Z., Xu, Y.: Slang detection and identification. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL) (pp. 881–889) (2019)
95. Wilson, S., Magdy, W., McGillivray, B., Garimella, K., Tyson, G.: Urban dictionary embeddings for slang NLP applications. In: Proceedings of the 12th Language Resources and Evaluation Conference (pp. 4764–4773) (2020)
96. Gopika, N.: Correlation based feature selection algorithm for machine learning. In: 2018 3rd International Conference on Communication and Electronics Systems (ICCES) (pp. 692–695). IEEE (2018). https://doi.org/10.1109/CESYS.2018.8723980
97. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**(3), 131–156 (1997). https://doi.org/10.1016/S1088-467X(97)00008-5
98. Koller, D., Sahami, M.: Toward Optimal Feature Selection. Stanford InfoLab, Stanford (1996)
99. Dadvar, M., Trieschnigg, D., Ordelman, R., de Jong, F.: Improving cyberbullying detection with user context. In: European conference on information retrieval (pp. 693–696). Springer, Berlin, Heidelberg (2013)
100. Zhong, H., Li, H., Squicciarini, A.C., Rajtmajer, S.M., Griffin, C., Miller, D.J., Caragea, C.: Content-driven detection of cyberbullying on the instagram social network. In: IJCAI (pp. 3952–3958) (2016)
101. Li, H.: Image Analysis of Cyberbullying using Machine Learning Techniques (2015)
102. Dadvar, M., Jong, F.D., Ordelman, R., Trieschnigg, D.: Improved cyberbullying detection using gender information. In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012). University of Ghent (2012)
103. Huang, Q., Singh, V.K., Atrey, P.K.: Cyber bullying detection using social and textual analysis. In: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia (pp. 3–6) (2014)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.